5448

0412.1
EF7.1

WAVE MECHANICS

ELEMENTARY THEORY

BY

J. FRENKEL

PROFESSOR AT THE PHYSICO-TECHNICAL
INSTITUTE, LENINGRAD

OXFORD
AT THE CLARENDON PRESS
1932

OXFORD UNIVERSITY PRESS
AMEN HOUSE, E.C. 4
LONDON EDINBURGH GLASGOW
LEIPZIG NEW YORK TORONTO
MELBOURNE CAPETOWN BOMBAY
CALCUTTA MADRAS SHANGHAI
HUMPHREY MILFORD
PUBLISHER TO THE
UNIVERSITY

PRINTED IN GREAT BRITAIN

536.32
F879

PREFACE

THE present book has a curious history which I would like briefly to narrate. Two years ago Dr. Urquhart of the University of Bristol started, with the approval of Professor Lennard Jones, a translation of my German book *Einführung in die Wellenmechanik*. This translation reached me at the beginning of 1931 while I was lecturing on Wave Mechanics at the University of Minnesota (Minneapolis, U.S.A.). It was intended that, after a slight revision and correction, it should be published by the Oxford Press. But during the three years that had elapsed since the appearance of the German original, my views on the subject had greatly developed, and in some respects (especially with regard to the analogy between matter and light) had been substantially modified. This maturing was largely due to the opportunity of presenting the subject to a highly qualified American audience with plenty of time for thinking it through.

As a result I became dissatisfied with my German version and started to write what is practically a new book, thus wholly undoing the work of Dr. Urquhart. I wish to record here how much I regret this, and also to thank Dr. Urquhart for the additional work he undertook of revising the English of my new version.

Now the large amount of leisure time at my disposal at Minneapolis—for which I feel greatly indebted to Professor H. A. Erikson, chairman of the Physics Department of the University of Minnesota—had the perhaps regrettable effect of enabling me to write much more than I initially intended. In this way it has come about that my relatively short and unpretentious *Einführung in die Wellenmechanik* will be replaced by a treatise in three volumes, each a work complete in itself, the first of which is now published. The second volume was practically completed in America and is in the press, while the third is planned for next year.

The general plan of the *Einführung in die Wellenmechanik* was preserved inasmuch as the first volume (replacing to some extent the first chapter of the German book) gives a general survey of the whole subject, or a first approach to it, using only elementary mathematics. As can be seen from the Contents, this first volume is fairly complete in itself, covering—although perhaps in a somewhat superficial way—the entire field of wave mechanics and of the new quantum

statistics connected with it. It contains also a number of applications to important and simple special problems.

The main object of the second volume is the development of the mathematical ideas which form the most general and most adequate expression of the physical principles embodied in the new mechanics. After a review of the correspondence between the new mechanics and classical mechanics, it will give a thorough presentation of the transformation theory and develop upon this basis the perturbation theory, with some illustrations. This will be followed by a relativistic refinement of the general theory, the investigation of the spin of the electron, and finally, by extensions to problems of many particles and of quantum electrodynamics.

The third volume will give a systematic treatment of various special problems connected with the structure of matter (complex atoms, molecules, crystals) and with the radiative and non-radiative interaction between material particles or bodies (collision phenomena, optical effects, and so on).

It is my pleasant duty to acknowledge my indebtedness to Professor E. Hill and Dr. G. Wetzel of the University of Minnesota, who have greatly helped me in the technical preparation of the manuscript, and to Professor R. H. Fowler of Trinity College, Cambridge, for his helpful criticism of a few passages and particularly for the readiness with which he met my expansive tendencies. The last chapter of the present volume, dealing with the electron theory of metals, was written in Leningrad after the proofs of the original manuscript were completed, so that there has been a delay of a few months in the appearance of the book.

LENINGRAD,
July 18th, 1932.

J. FRENKEL.

CONTENTS

I. LIGHT

1. Corpuscular and Wave Theories of Light 1
2. The Electromagnetic Theory of Light and the Theory of Relativity 6
3. Einstein's Relativity Mechanics and the Light Quantum Hypothesis 9
4. The Relativistic Formulation of the Wave-Corpuscle Parallelism . 14

II. MATTER

5. Extension of the Wave-Corpuscle Parallelism to Matter . . 19
6. Interference and Diffraction of Cathode Rays 23
7. Group Velocity of Wave Processes and the Theory of Wave Packets 27
8. The Probability Concept 32
9. Wave Packets as Probability Packets; Uncertainty Relations . 38
10. Heisenberg's Theory of the Uncertainty Relation and the Role of the Observer 47

III. WAVE MECHANICS OF THE MOTION OF A PARTICLE IN A FIELD OF FORCE

11. Bohr's Quantum Theory of the Atom and Standing Cathode Waves 53
12. Reflection and Transmission Phenomena 65
13. The Wave Equation of Schrödinger and its Application to the Harmonic Oscillator 74
14. Wave Mechanics of a Hydrogen-like Atom; Comparison with Isotropic Oscillator and Rotator 84
15. The Method of the 'Potential Staircase' for the Approximate Solution of One-dimensional Wave-mechanical Problems . 99
16. Application of the Preceding Method to some Simple Problems . 107
17. Superposition of Different Quantized States: Wave-mechanical Theory of Light Emission 124
18. Wave-mechanical Theory of Transition Processes 137
19. Vector Waves: Polarization of Light and Electron Spin . . 149

IV. WAVE MECHANICS OF A SYSTEM OF PARTICLES

20. Method of Copies: Theory of Quantized Waves . . . 156
21. The Method of Configuration Space 166
22. The Symmetry Properties of Functions representing a System of Identical Particles 172
23. The Pauli Exclusion Principle and its Consequences . . 181
24. Relation between Matter and Light 187

V. STATISTICAL MECHANICS

25. Statistical Equilibrium of a Gas at the Zero Point of the Temperature 192
26. General Law of Statistical Equilibrium 194
27. Application of the General Law of Statistical Equilibrium to an Ideal Gas 197
28. Alternative Derivation of the Equilibrium Distribution of a Gas . 206
29. Limiting Forms of the Quantum Distribution Laws . . 210

VI. APPLICATION OF THE QUANTUM STATISTICS TO THE ELECTRON THEORY OF METALS, HEAT MOTION, AND RADIATION

30. The Theory of an Electron Gas; Free Electrons in Metals . . 214
31. Refinement and Extension of the Electron Gas Theory . . 225
32. The Calculation of the Self-consistent Field in a Metal and on its Surface 233
33. Electric Properties of Contacts between Conducting Bodies . 241
34. Electrical and Thermal Conductivity of Metals (Corpuscular Theory) 251
35. Wave-mechanical Theory of Collisions in an Electron Gas . 257
36. 'Collectivistic' Treatment of the Motion of Electrons in a Metal . 263
37. The Theory of a 'Phonon Gas' (Mechanical Heat in a Solid) and of its Interaction with the Electron Gas 266
38. The Theory of a Photon Gas (Thermal Radiation) . . . 272

I

LIGHT

1. Corpuscular and Wave Theories of Light

The development of our ideas on the nature of light shows a remarkable alternation between the corpuscular conception on the one hand and the wave conception on the other. We usually regard Newton as the founder of the corpuscular theory of light and Huygens as the founder of the wave theory, although as a matter of fact both theories are much older.

It is worthy of remark that Newton himself, in his optical considerations, often used the wave theory. However, he always preferred the corpuscular theory.

Because of the great scientific authority of Newton, the corpuscular theory of light dominated the known field of optical phenomena until the beginning of the nineteenth century. Besides the discovery of the diffraction of light which led to the new wave optics, we must mention the famous experiment of Fizeau which, in the old struggle between the corpuscular and the wave theory of light in the narrow region of geometrical optics where both up to that time had appeared as equivalent, decided in favour of the latter.

Because of its significance in the new theory of material waves we shall consider, in some detail, the theoretical foundation of Fizeau's experiment.

According to the corpuscular theory of Newton, we must regard light rays as the paths of definite corpuscles of light or 'atoms of light'. These corpuscles are emitted in all directions from the glowing body and move in empty space or in a homogeneous material medium with uniform motion in a straight line, i.e. just like ordinary material particles in the absence of external forces or mutual interactions. Reflection and refraction of light rays at the surface between two different homogeneous bodies, e.g. air and water, are explained, on the corpuscular theory, by the action of definite forces active in the surface transition layer. These 'light forces' must act perpendicular to the surface and produce an alteration in the normal component of the velocity without changing the tangential component. If they are directed from the second medium into the first, they can either stop the particles before they penetrate through the transition layer into the second medium and, by throwing them back into the first medium,

5443

produce a *reflection* of light, or they can reduce the normal component of the velocity, allowing the particles, however, to penetrate into the second medium. This *transmission* of light is in general accompanied by a change in direction of the light rays, i.e. by refraction, the rays being bent away from the perpendicular to the surface. The same phenomenon of transmission and refraction occurs if the forces in the transition layer are directed from the first medium into the second, so that they produce an increase in the normal component of the velocity. In this case, however, the rays are bent in the direction of the perpendicular to the surface.

Let the velocity of a corpuscle in the first medium be $OC = v$ and in the second $OC' = v'$, i.e. we shall suppose that the light is transmitted (Fig. 1). Let the common tangential component be $OA = OA'$ and the normal components OB and OB'.

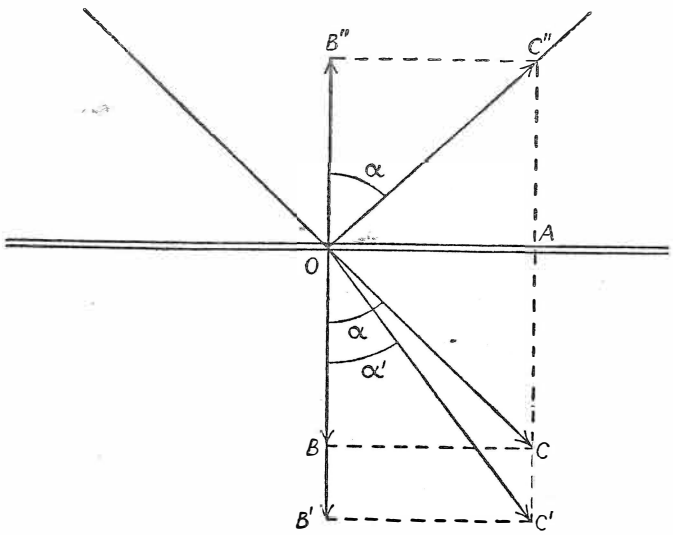

FIG. 1.

Then $$OA = v \sin \alpha = v' \sin \alpha',$$

where α and α' are the angles of incidence and refraction respectively.

Hence $$\frac{\sin \alpha}{\sin \alpha'} = \frac{v'}{v}. \qquad (1)$$

This formula expresses the well-known Snell's Law if we can assume that the velocity v' is a quantity independent of the direction of incidence—a quantity which is just as characteristic of the second medium as v is of the first. Now it is easy to see that this condition must, in fact, be fulfilled and can be regarded as an immediate con-

sequence of the conservation of energy. If U denotes the work done by the 'light forces' when a corpuscle goes from the first into the second medium, and if, further, we assume that the motion of the corpuscles obeys the usual laws of classical (Newtonian) mechanics, then

$$\tfrac{1}{2}m(v')^2 = \tfrac{1}{2}mv^2 + U,$$

(m = mass of corpuscle), i.e.

$$v' = \sqrt{v^2 + \frac{2U}{m}} = \text{const.}$$

The quantity U, which is assumed to be constant, gives the difference of the potential energy of the corpuscle of light in the first and in the second medium.

If we denote the refractive index, i.e. the ratio $\frac{\sin\alpha}{\sin\alpha'}$, by μ, then by (1) we get for it the following 'dynamical' expression:

$$\mu = \sqrt{1 + \frac{U}{\frac{1}{2}mv^2}}. \tag{2}$$

If U is negative and if in absolute magnitude it is larger than $\frac{1}{2}m(v\cos\alpha)^2$ —the part of the kinetic energy of a particle (in the first medium) corresponding to the normal component of its velocity—then the particle will be thrown back into the first medium with its normal velocity component reversed (OB'') and we shall get the well-known reflection law: the direction of the reflected ray OC'' makes with the normal to the surface the same angle α as that of the incident ray.

We shall now consider how the laws of refraction and reflection are explained in terms of the wave theory of light.

The conception of light rays, or better still, of a bundle of light rays, is here replaced by the idea of a train of waves. Usually we identify the velocity of propagation of the waves with the velocity of motion of the corpuscles of light. We shall, however, make no such assumption but shall denote the wave velocities in the first and the second medium by new letters, namely, w and w'. In Fig. 2 these velocities are represented by the lines PN and MQ' which give the displacement of the refracted wave front in the respective media in unit time. The line of refraction (represented in the figure by the point M or N) thus moves along the surface with the velocity

$$MN = \frac{w}{\sin\alpha} = \frac{w'}{\sin\alpha'}.$$

Hence

$$\frac{\sin\alpha}{\sin\alpha'} = \frac{w}{w'} = \text{const.} \tag{3}$$

This formula agrees with formula (1) in predicting that the refractive index does not depend upon the direction of incidence. It gives, however, for μ an expression which appears to be exactly the reciprocal of (1), or rather would so appear if we identified the wave velocities w and w' with the corresponding corpuscular velocities v and v'.

Replacing the refracted wave fronts MP', NQ' in the second medium by the reflected wave fronts MP'' and NQ'' in the first medium, we obtain in a similar way the old law of reflection.

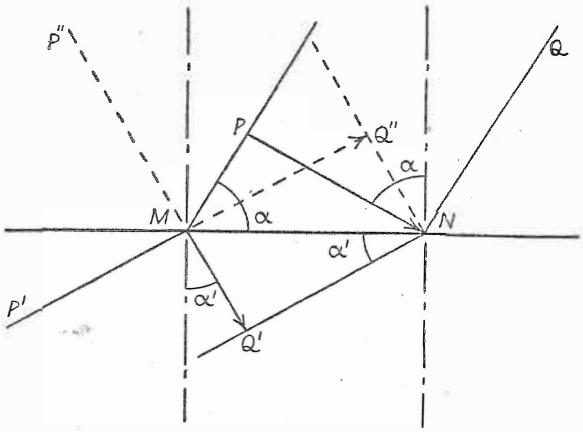

FIG. 2.

When light rays travel from air into water we have $\alpha > \alpha'$, i.e. $\mu > 1$. According to the corpuscular theory of Newton it would then follow that $v' > v$. Now Fizeau actually measured the velocity of light in water and found that it is *smaller* than the velocity of light in air. The ratio of the two velocities agreed with the prediction of the wave theory, i.e. with the conception of the velocity of light as a wave velocity according to formula (3).

By this 'crucial experiment' alone the struggle between the corpuscular and the wave theories of light was regarded as settled. In addition there arose a series of newly discovered interference and diffraction phenomena which were quite incomprehensible from the corpuscular point of view but which were very easily explained by the wave theory, even to the last quantitative detail.

Independently, however, of these new facts, the corpuscular and wave theories of reflection and refraction could be considered as equally satisfactory only in so far as the *direction* of the reflected and transmitted rays was concerned. There is a fundamental difference between them when we consider the relative *intensity* of these rays. According to the

corpuscular theory we should expect either complete reflection or complete transmission, because all the corpuscles are supposed to move under identical conditions and therefore should all experience the same fate, i.e. either *all* be thrown back into the first medium, or *all* pass into the second.† Actually total reflection takes place only when $\mu < 1$ and $\sin\alpha > \mu$, which would imply $\sin\alpha' > 1$. This agrees with the corpuscular theory, for in this case the kinetic energy corresponding to the normal component of the velocity is less than the work, $-U$, to be done against the forces acting in the transition layer. There is, however, a *partial* reflection of light even if $\sin\alpha < \mu$ when according to the corpuscular theory there should be no reflection at all, but complete transmission. To reconcile the corpuscular theory of light with the facts we must therefore abandon the idea that identical causes produce identical effects, that is, we must forgo the idea of *determinism* in corpuscular phenomena. As we shall see later, this conclusion has actually been forced on physicists in recent times in connexion with similar phenomena displayed by cathode rays. At the beginning of the nineteenth century such a conclusion was by no means necessary. It was far more reasonable to consider the failure of the corpuscular theory to explain the phenomenon of *partial* reflection and transmission of light as a conclusive proof against it and in favour of the wave theory, which provides for this phenomenon a very natural and simple explanation. Huygens was the first to give such an explanation, though it was only qualitative. He based it on the same two principles which he used to explain the geometrical aspects of reflection and refraction. His first principle is that each point of a wave front can be considered as a centre of an elementary spherical wave. His second principle is that these elementary spherical waves combine with one another (principle of interference), thus producing the displaced wave front at the next instant. It followed from this that a disturbance reaching the surface of separation between two media will be propagated in general both into the second and also back into the first medium.

In the hands of Fresnel this idea of Huygens was transformed into a quantitative theory which allowed one to calculate the relative intensity of the reflected and the transmitted rays for various angles of incidence, in full agreement with experimental data.

† Newton tried to escape from this difficulty by assuming that the light particles could have 'fits of reflection' or 'fits of transmission'—an idea very similar to the present views on the question, the conception of 'fits' being now replaced by that of probabilities (see below).

By the middle of the nineteenth century the wave theory of light—mainly because of the important contributions of Fresnel—had attained such a beautiful form, a form both self-consistent and inclusive of the whole region of macroscopic optical phenomena known at that time, that the possibility of an essential revision of the theory in the sense of a return to a corpuscular theory of light appeared to be completely excluded.

2. The Electromagnetic Theory of Light and the Theory of Relativity

The electromagnetic theory of light, which was introduced by Maxwell in the second half of the last century, retained the formal foundations of Fresnel's theory and only altered and made more definite the physical meaning of the quantities concerned, as well as our conception of the nature of light waves. It showed that these waves are identical with Maxwell's electromagnetic waves. The latter were realized experimentally by Hertz, and in recent years have found an extensive technical application in wireless telegraphy and telephony. The difference between light and radio waves consists, as is well known, only in their wave-lengths.

Up to the end of the nineteenth century electromagnetic waves were regarded as mechanical vibrations in the 'luminiferous ether'—a conception also introduced by Huygens. Einstein's theory of relativity showed, however, that a mechanical explanation of electromagnetic waves is neither necessary nor possible. Thus light waves were robbed of their material basis and were transformed into periodic vibrations of an electromagnetic field in empty space. Their connexion with material bodies is limited to the acts of emission and absorption, which correspond respectively to the sending and receiving of radio waves. Here we have single particles—atoms or molecules—serving as transmitting aerials and receiving stations.

According to the theory of relativity light waves are always propagated with the same true velocity $c = 3 \times 10^{10}$ cm./sec. In material bodies the velocity of propagation is different from this and must be regarded as an *apparent* velocity produced as follows:

A material body is not a continuous medium but a system of particles (electrons, nuclei) practically without extension, situated in empty space. The electromagnetic elementary waves emitted by any particle are propagated with the true light velocity c just as the primary (incident) waves which produce the forced vibrations of this particle. Through superposition or interference of the elementary waves, there arises a complicated wave form which, from the macroscopic stand-

point, can be regarded as a system of ordinary waves that move with a velocity w different from c. The weaker the binding of the electrons of the body, or, in general, the easier they can be brought into vibration (thanks to resonance for example), the larger is the amplitude of the secondary light waves emitted by them, and the greater must the apparent velocity w differ from c. The 'displaceability' or capability of vibration of the electrons in the atoms of any body is, in electrical theory, described by the dielectric constant ϵ of the body (for a body with rigidly bound electrons—just as for empty space—ϵ would have the value 1). From this it follows that the ratio c/w which determines the refractive index μ of the body with respect to free space must differ by a greater amount from unity, the greater the dielectric constant. In fact the two quantities satisfy the relation discovered by Maxwell:

$$\mu = \sqrt{\epsilon}.$$

Through superposition of the primary and secondary elementary waves all peculiarities which distinguish light propagation in material bodies from that in empty space (among others, light absorption) can be explained. A detailed discussion of this question would, however, lie outside the scope of this book. What alone is essential for us is that light waves of definite origin, i.e. elementary waves, are always propagated with the same velocity c, irrespective of whether they are travelling in macroscopic empty space, e.g. between two distant stars, or in microscopic empty space, i.e. between two atoms, or more exactly between the electrons of an ordinary material body.

According to the theory of relativity this *true* velocity of light differs from all other velocities in that its magnitude remains unaltered when a change is made from a given coordinate system to any other which moves with uniform translational velocity relative to the first. This property can be considered as a fact of experience known from Michelson's experiment and a series of other experiments of the same kind. One can, however, also consider it as a direct consequence of the principle of relativity, i.e. of the principle of the complete equivalence of different coordinate systems moving relatively to one another with uniform translational motion due to the fundamental impossibility of determining which of these systems is actually at rest in empty space.

The principle of the relativity of velocity (in the above sense) together with the invariance principle of the true velocity of light arising therefrom, form, as we know, the theoretical foundations of Einstein's (special) theory of relativity.

The content of this theory can most simply be summarized as a certain—although not complete—equivalence between the three space dimensions and the time dimension, an equivalence which is expressed mathematically by the invariance of the square of the space-time interval

$$s^2 = (x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2-c^2(t_2-t_1)^2. \qquad (4)$$

Here x_1, y_1, z_1 are the rectangular coordinates of the point where an event (1) takes place at the instant t_1; the quantities x_2, y_2, z_2, t_2 have the same meaning for a second event (2). The invariance holds for two coordinate systems moving relatively to one another with uniform translational velocity. One has, therefore, for two such systems independent of the relative velocity v, $s = s'$, where

$$(s')^2 = (x_2'-x_1')^2+(y_2'-y_1')^2+(z_2'-z_1')^2-c^2(t_2'-t_1')^2.$$

If $v = 0$, i.e. if the systems are at rest relatively to one another, then $t' = t$, while the space coordinates x, y, z and x', y', z' are connected with one another by linear relations whose coefficients depend on the relative orientations of the two coordinate systems. If $v \neq 0$ we have, in general, relations of the same kind between the quantities x, y, z, t and x', y', z', t'. These relations assume a completely symmetrical form (a linear orthogonal transformation) if we replace the time t by the imaginary quantity $l = \sqrt{(-1)}ct = ict$ proportional to it, and t' by $l' = ict'$. The i, which is necessary because of the negative sign of the time term in the expression (4), limits the physical equivalence of space and time. Formally, however, both can be fused into a four-dimensional continuum in which all directions are equivalent, just as are the different directions in ordinary three-dimensional space. The rotation of the time axis through an imaginary angle (and the corresponding rotation of one or all of the space axes with preservation of their orthogonality) corresponds to a transition from the original spatial coordinate system to another which moves relatively to the first with a velocity v proportional to the angle of rotation.

In the simplest case where both space systems are similarly oriented and the relative motion takes place in the direction of the x-axis, the relations assume the following well-known form (Lorentz transformation):†

$$x' = \frac{x-vt}{\sqrt{(1-v^2/c^2)}}, \qquad y' = y, \qquad z' = z, \qquad t' = \frac{t-xv/c^2}{\sqrt{(1-v^2/c^2)}}. \qquad (5)$$

† We obtain these formulae from the ordinary transformation formulae of analytical geometry $x' = x\cos\alpha - l\sin\alpha$, $l' = x\sin\alpha + l\cos\alpha$ for a rotation in the (x, l) plane. If we put $\tan\alpha = v/ic$ we get: $x' = \cos\alpha(x-vt)$ and $l' = \cos\alpha(l+xv/ic)$ or $t' = \cos\alpha(t-xv/c^2)$, where $\cos\alpha = (1+\tan^2\alpha)^{-\frac{1}{2}} = (1-v^2/c^2)^{-\frac{1}{2}}$.

From the union of space and time there follows a similar union of all ordinary three-dimensional vector quantities with the corresponding scalars (i.e. quantities which are independent of the orientation of the spatial system of reference) to form four-dimensional vectors. These four-dimensional vectors can be represented graphically by space-time distance vectors with components x, y, z, t (or l). The three-dimensional vectors are then given by their space projections and the corresponding scalars by their time projections.

In order that the physical laws may remain identical for observers who belong to different systems of reference which move relatively to one another, these laws must be formulated by *equating* the corresponding components of various four-dimensional vector quantities. Then such equations, if they hold for one definite coordinate system, must hold for any other coordinate system, because all components have the same transformation properties.

We cannot go into the details here. We shall only remark that the fundamental equations of the electromagnetic theory of light (for empty space) satisfy this invariance or covariance condition, while the equations of Newtonian mechanics, and consequently the Newtonian corpuscular theory of light, contradict it. The remodelling of classical mechanics on relativity lines, which Einstein undertook in 1905, suggested the corresponding remodelling as well of the old corpuscular theory of light, the more so because the invariant true velocity of light plays a fundamental role in all mechanical relations. Besides, this remodelling was demanded by new experiments which, perhaps by chance, were carried out at about the same time and which led to the conception of the quantum theory (Planck's radiation formula, photoelectric effect). Einstein had the courage, in spite of the brilliant success of the wave theory during the previous century, to undertake such a remodelling. This led him, as we know, to the formulation of the corpuscular theory of light in a new, improved, and extended form.

3. Einstein's Relativity Mechanics and the Light Quantum Hypothesis

In Einstein's mechanics of a material particle the time of observation, as well as the space coordinates, are treated as dependent variables. The role of the time as independent variable is taken over by the invariant 'proper time' τ whose differential is defined by the formula

$$d\tau = \frac{ds}{ic} = \frac{1}{ic}\sqrt{(dx^2+dy^2+dz^2-c^2dt^2)}. \qquad (6)$$

This differential is therefore equal to the invariant space-time distance ds between two neighbouring 'events' divided by ic. If we denote the velocity of the particle

$$\sqrt{\left\{\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2+\left(\frac{dz}{dt}\right)^2\right\}}$$

by v, we can write $d\tau$ in the form

$$d\tau = dt\sqrt{\left(1-\frac{v^2}{c^2}\right)}. \tag{6a}$$

This formula shows that the proper time reduces practically to the ordinary time in the limiting case of $v \ll c$.

Einstein's equations of motion are

$$m_0\frac{d^2x}{d\tau^2} = F_x, \quad m_0\frac{d^2y}{d\tau^2} = F_y, \quad m_0\frac{d^2z}{d\tau^2} = F_z, \quad m_0\frac{d^2l}{d\tau^2} = F_l, \tag{7}$$

where F_x, F_y, F_z, F_l are the components of a four-vector which we can define as the 'four-dimensional force' and where m_0 corresponds to the ordinary mass.

If we introduce as independent variable the ordinary time instead of the proper time, we get, according to (6a):

$$\frac{d}{dt}\frac{m_0 v_x}{\sqrt{\left(1-\frac{v^2}{c^2}\right)}} = F_x\frac{d\tau}{dt}, \quad \frac{d}{dt}\frac{m_0 v_y}{\sqrt{\left(1-\frac{v^2}{c^2}\right)}} = F_y\frac{d\tau}{dt}, \quad \frac{d}{dt}\frac{m_0 v_z}{\sqrt{\left(1-\frac{v^2}{c^2}\right)}} = F_z\frac{d\tau}{dt}, \tag{7a}$$

and

$$\frac{d}{dt}\frac{m_0 ic}{\sqrt{\left(1-\frac{v^2}{c^2}\right)}} = F_l\frac{d\tau}{dt}. \tag{7b}$$

Equations (7a) can be summarized in the vectorial form

$$\frac{d}{dt}\mathbf{g} = \mathbf{f}, \tag{8}$$

where the vector

$$\mathbf{g} = \frac{m_0\mathbf{v}}{\sqrt{\left(1-\frac{v^2}{c^2}\right)}} \tag{8a}$$

is to be considered as the momentum of the particle, and the vector

$$\mathbf{f} = \mathbf{F}\frac{d\tau}{dt} = \mathbf{F}\sqrt{\left(1-\frac{v^2}{c^2}\right)}$$

as the ordinary three-dimensional force. From this it follows that the vector $\mathbf{F}$ is the impulse of this force per unit of proper time.

Equation (7b) is not independent of (7a). One has in fact from (6)

$$\left(\frac{dx}{d\tau}\right)^2+\left(\frac{dy}{d\tau}\right)^2+\left(\frac{dz}{d\tau}\right)^2+\left(\frac{dl}{d\tau}\right)^2=-c^2,$$

and hence

$$\frac{dx}{d\tau}\frac{d^2x}{d\tau^2}+\frac{dy}{d\tau}\frac{d^2y}{d\tau^2}+\frac{dz}{d\tau}\frac{d^2z}{d\tau^2}+\frac{dl}{d\tau}\frac{d^2l}{d\tau^2}=0,$$

or from (7)

$$F_x\frac{dx}{d\tau}+F_y\frac{dy}{d\tau}+F_z\frac{dz}{d\tau}+F_l\frac{dl}{d\tau}=0.$$

We have therefore

$$-icF_l=F_x\frac{dx}{dt}+F_y\frac{dy}{dt}+F_z\frac{dz}{dt}=f_x\frac{dx}{d\tau}+f_y\frac{dy}{d\tau}+f_z\frac{dz}{d\tau}=\mathbf{f}\cdot\mathbf{v}\frac{dt}{d\tau}.$$

From this it follows that the quantity $-icF_l$ gives the work of the force $\mathbf{f}$ per unit of proper time. Equation (7b) can therefore be written in the form

$$\frac{d}{dt}\frac{m_0c^2}{\sqrt{(1-v^2/c^2)}}=\mathbf{f}\cdot\mathbf{v}, \tag{9}$$

and the quantity

$$\epsilon=\frac{m_0c^2}{\sqrt{(1-v^2/c^2)}} \tag{9a}$$

may be interpreted as the energy of the particle. This energy is compounded from the kinetic energy $m_0c^2\left(\frac{1}{\sqrt{(1-v^2/c^2)}}-1\right)$ which, for small values of v/c, assumes the usual form $\frac{1}{2}m_0v^2$ and the 'rest energy' m_0c^2 which is proportional to the 'rest mass' m_0 of the particle. One usually defines the quantity

$$m=\frac{m_0}{\sqrt{(1-v^2/c^2)}} \tag{10}$$

as the mass of the moving particle. With this definition we have from (8a) and (9a)

$$\mathfrak{g}=m\mathbf{v}, \qquad \epsilon=mc^2. \tag{10a}$$

It is most important to observe, firstly that in relativity mechanics the momentum or impulse $\mathfrak{g}$ and the energy ϵ combine to form the space and time projections respectively of a four-dimensional 'momentum-energy vector' (it would be more correct to define the time projection by $i\epsilon/c=imc$), and secondly the peculiar dependence of the mass m on the velocity of motion v as determined by (10).

From this equation it follows that v cannot exceed the critical value c. The velocity of light c plays, therefore, in Einstein's mechanics, the role of a limiting velocity which can only be reached by a particle with *vanishing rest mass*. The case $m_0=0$ has, in Newtonian mechanics,

absolutely no sense. In Einstein's mechanics, on the other hand, it represents those particles which move with the velocity of light—assuming, of course, that such particles actually exist.

If we postulate the existence of these particles we must obviously conceive them to be corpuscles of light, and we are led in this way quite naturally to the corpuscular theory of light. As already remarked, Einstein's new corpuscular theory of light came just at the right time to explain some new results of experimental and theoretical research and to predict many other quantum phenomena. On the other hand, interference and diffraction of light, which were so simply explained by the wave theory, remained, from the new standpoint, completely unintelligible.

The main difference between Einstein's theory of light and that of Newton is that in the former the velocity of the light corpuscles, not only in empty space, but also in material bodies (which must be regarded as vacua containing point-like material particles), has always the same value c. Einstein's explanation of the refraction of light at the surface between two different bodies must, therefore, be quite different from that of Newton. It must involve absorption and re-emission of the light corpuscles by the atoms of the body—corresponding to the superposition of the primary and the secondary elementary waves in the electromagnetic theory of light.

Moreover, Einstein's light corpuscles or 'light-quanta', or 'photons' according to the now current terminology, exist only when they move with the velocity c. Since they possess no material substance such as can be specified by some value of the rest mass different from zero they vanish at the instant they come to rest (or at the moment when their velocity becomes smaller than c). This occurs, for example, when a photon is absorbed by a material atom. The light quantum does not stick to the atom like a light corpuscle of Newton, but is 'digested' by the atom as it were and so ceases to exist as a particle, its energy ϵ and momentum $\mathfrak{g}$ being both transferred to the absorbing atom.

According to (10a) we have

$$\mathfrak{g} = \epsilon \frac{\mathbf{v}}{c^2}. \tag{11}$$

In the limiting case $v = c$ we must have, between the momentum and the energy, the relation

$$g = \epsilon/c. \tag{11a}$$

Except for this relation, the two quantities remain undetermined since

by (8a) and (9a) when $m_0 = 0$ and $v = c$ the ratio $m_0/\sqrt{(1-v^2/c^2)}$ assumes the form 0/0.

Besides energy and momentum, a mass $m = \epsilon/c^2 = g/c$ can also be assigned to Einstein's photons. Of these three properties one usually considers the energy ϵ as the 'primary' and most essential. One is accustomed, therefore, to regard light quanta as 'energy quanta'. This, however, is not fundamentally justified since the momentum is just as essential as the energy and should not be separated from it. The preference given to the energy at the expense of the momentum can be understood, however, if we consider the origin of Einstein's light quantum hypothesis.

This hypothesis in fact was not developed from the theory of relativity alone, the way having been prepared for it by Planck's work on the theory of heat radiation. To explain the observed intensity distribution in the spectrum of this radiation, Planck, as is well known, had to assume that radiation energy is not continuously absorbed or emitted by atoms, but in a step-wise manner in quite definite finite amounts or 'energy quanta'. Planck did not connect these energy quanta with a discontinuous corpuscular structure of the radiation. He considered that they were absorbed and emitted in the usual way corresponding to the electromagnetic theory of light. The absorbing and emitting atom he considered as a harmonic oscillator with a definite vibration frequency ν which must coincide with the frequency of the absorbed or emitted electromagnetic waves. Planck then found from thermodynamic considerations that the magnitude of his energy quantum is connected with the frequency by the relation

$$\epsilon = h\nu, \tag{12}$$

where $h = 6{\cdot}55 \times 10^{-27}$ erg. sec. is the constant named after him.

Five years later Einstein, guided by the theory of relativity, endeavoured to make Planck's theory still more revolutionary by introducing the conception that Planck's energy quanta are expelled from the atoms not in the shape of spherical waves, but as quasi-material corpuscles with a definite direction of motion. To these light corpuscles or photons he ascribed, besides energy, also definite momenta determined by (11a). Hence, because of the conservation of momentum, an atom, when it radiates, must undergo a recoil—just as a gun does when fired.

In collision with another atom (or electron) the photon would either be reflected (scattered) or absorbed. In the latter case it would transfer

to the atom, besides its energy, also its momentum, a transfer which gives a simple description of the phenomenon of light pressure.

These bold ideas of Einstein soon received a series of brilliant experimental confirmations, of which the photo-electric effect and the Compton effect are striking examples. We shall not consider more closely these well-known phenomena but shall merely emphasize the following: In the application of Einstein's light quantum theory to these and other phenomena nothing was said of the mechanism of the mutual action between the atoms (or electrons) and the photons. There was no definite conception of the process of the emission and absorption of a photon, or of its scattering by a free electron (in the Compton effect). To describe these phenomena in the usual classical sense, we would have to introduce definite mutual interactions between material particles and photons and follow the *course in time* of the processes under consideration by means of the relativity equations of motion. This was, however, never carried out. Rather, one treated these processes as '*momentary*' or better, as 'complete' acts (as for example the collision processes of elementary mechanics), and was satisfied with determining the final result by means of the laws of the conservation of energy and momentum for the total system (atom or electron and photon) on the one hand and Planck's relation (12) between energy and frequency on the other.

This relation has, in Einstein's theory, a quite new and purely symbolic meaning. We are not, like Planck, dealing with energy quanta which are radiated in the form of waves, but with the energies and momenta of individual quasi-material corpuscles. The idea of a vibration frequency here seems to be completely unsuitable. However, this idea is unavoidable since we can only determine the energy of an Einstein light quantum by a determination of the wave-length of the light under consideration by means of some interference or diffraction experiment.

4. The Relativistic Formulation of the Wave-Corpuscle Parallelism

During the first stage of the struggle between the wave and the corpuscular theories of light—between Newton and Huygens—each point of view appeared in the form of an alternative; they were mutually exclusive. The new stage of this struggle—the stage connected with the name of Einstein—is distinguished by a quite different relation between the two theories and, indeed, by a peculiar mutual union of the two theories, each supplementing the other. One series of optical

phenomena such as interference and diffraction of light can only be explained by the wave theory. Another series of phenomena such as the photo-electric effect and the Compton effect require the corpuscular theory to interpret them. The same light which in the processes of the first group reveals itself as monochromatic waves with a frequency ν, appears in the processes of the second group as homogeneous corpuscular radiation in which the energy ϵ of each particle is equal to $h\nu$.

The frequency of light waves is, as we know, not directly measured, but is calculated from the wave-length which, in its turn, is determined by means of interference experiments; the frequency then follows from the formula $\nu = c/\lambda$. If we compare this relation with the relation $\epsilon = gc$ between the energy and the momentum of a light quantum, we obtain, besides the Planck formula $\epsilon = h\nu$, also the following:

$$g = h/\lambda. \tag{13}$$

The wave-length λ, which we could define as the spatial period of vibration, corresponds to the time vibration period $\tau = 1/\nu$. The vibration frequency ν, i.e. the number of vibrations in unit time, corresponds to the so-called *wave number* $k = 1/\lambda$, i.e. the number of waves in unit length (for a given instant t). Since g is the magnitude of a (three-dimensional) vector $\mathfrak{g}$, we are led to consider k as the magnitude of a similarly directed vector which determines the direction of propagation of the light waves and to replace the relation (13) by the more complete relations

$$g_x = hk_x, \qquad g_y = hk_y, \qquad g_z = hk_z. \tag{13a}$$

In this way we emphasize the relation between the corpuscular and the wave theory of light in that the direction of motion of the light quanta or photons at any point is the same as that of the light waves.

The Planck relation $\epsilon = h\nu$ appears at first to be fundamentally unfounded. If we accept the dual nature of light as a fact of experience, the question then arises: why must the energy of the light corpuscles be directly proportional to the vibration frequency of the waves and not, for example, to the wave-length or to any other more complicated function of ν or λ?

This question can be satisfactorily answered on relativity grounds. One can easily show that the relation of the corpuscular and the wave quantities defined by the formulae (12) and (13a) is the only one which satisfies the invariance condition of the theory of relativity.

For this purpose we consider the equation of a system of plane sine waves with wave-length λ, frequency ν, and a direction of propagation

which makes angles α, β, γ with the coordinate axes. This equation has the form

$$\psi = \psi_0 \cos 2\pi\left(\frac{x\cos\alpha + y\cos\beta + z\cos\gamma}{\lambda} - \nu t\right), \tag{14}$$

where ψ is the vibrating quantity and ψ_0 its amplitude. If we introduce the 'wave vector' $\mathbf{k}$ with components

$$k_x = \frac{1}{\lambda}\cos\alpha, \qquad k_y = \frac{1}{\lambda}\cos\beta, \qquad k_z = \frac{1}{\lambda}\cos\gamma,$$

then

$$\psi = \psi_0 \cos 2\pi(k_x x + k_y y + k_z z - \nu t). \tag{14a}$$

The vibration phase $\phi = 2\pi(k_x x + k_y y + k_z z - \nu t)$ must, because of its physical meaning (dimensionless character), represent an invariant quantity, i.e. a pure four-dimensional scalar. Since, however, x, y, z, ict are the components of a four-vector, the quantities k_x, k_y, k_z, $i\nu/c$ must also be the components of a four-vector, for only under this condition does the sum $xk_x + yk_y + zk_z - t\nu$ transform like the sum

$$x^2 + y^2 + z^2 + (ict)^2.$$

The latter, however, remains invariant when we transform from the original coordinate system to a 'dashed' (moving) system. Consequently the former must also transform into the similar and equivalent sum $x'k'_{x'} + y'k'_{y'} + z'k'_{z'} - t'\nu'$. The phase of a plane harmonic wave train is therefore characterized by a four-vector

$$k_x, \qquad k_y, \qquad k_z, \qquad i\nu/c.$$

On the other hand we have seen in § 2 that the quantities

$$mv_x = g_x, \qquad mv_y = g_y, \qquad mv_z = g_z, \qquad imc = i\epsilon/c$$

form a four-vector which determines the momentum and the energy of a point-like particle (compare (7a) and (7b)). If, therefore, there exists in general a relation between these corpuscular theory quantities and the above wave theory quantities, it can only have the form

$$\frac{g_x}{k_x} = \frac{g_y}{k_y} = \frac{g_z}{k_z} = \frac{\epsilon}{\nu} = h. \tag{15}$$

Otherwise it would depend on the choice of the coordinate system, which is obviously impossible.

The ratio of the mechanical and optical quantities indicated in (15) by h could in principle, however, be an arbitrary function of the invariant quantities

$$k_x^2 + k_y^2 + k_z^2 - \frac{\nu^2}{c^2} \quad \text{and} \quad g_x^2 + g_y^2 + g_z^2 - \frac{\epsilon^2}{c^2}.$$

From the formulae $\nu = c/\lambda$ and $\epsilon = gc$ it follows that both these in-

variants are zero. The quantity h must therefore be a constant. Thus we obtain our uniqueness proof.

Had Einstein developed his theory of relativity and created the new light quantum theory before Planck's work on the quantum theory, he would have been able to predict theoretically Planck's relation between energy and frequency.

The connexion between light quanta and light waves is not yet exhausted by the relations (15). These relations contain only such quantities as, according to both conceptions, determine the *quality* of the light and not its *quantity*. According to the corpuscular theory the 'quantity' or 'intensity' of the light must be determined by the number of corpuscles which cross a unit surface perpendicular to their direction of motion (direction of the light rays) in unit time. On the other hand, in the wave theory the light intensity is determined by the square of the amplitude of the vibrations at the point considered. The correspondence between the two conceptions with regard to intensity can be formulated thus:

$$\text{(Number of corpuscles) } N \sim \psi_0^2 \text{ (square of the amplitude).} \qquad (16)$$

Here we must understand by N the *average* number of corpuscles in unit volume (i.e. the 'density' or 'concentration' of the photons). The number of corpuscles which cross a unit surface perpendicular to their direction of motion in unit time is simply equal to the product of N by the velocity of light c. This product we can consider as the magnitude of a vector $\mathbf{J}$ which we shall call the 'current density'. The products of its components J_x, J_y, J_z with the energy ϵ, or with the components of the momentum g_x, g_y, g_z of a single corpuscle, then give a measure of the energy or the momentum respectively which is transferred by the light rays in unit time across unit surface perpendicular to one of the coordinate axes.

According to the theory of relativity we must add to the three space projections of the current density still a (fourth) time projection J_t. This can be defined, however, as can easily be seen, simply as the number N introduced above. Thus for the complete determination of the intensity or 'quantity' or light according to the corpuscular theory we have sixteen quantities which are given by multiplication of the components of the four-dimensional vectors J_x, J_y, J_z, J_t and g_x, g_y, g_z, ϵ and which can be regarded as forming a four-dimensional *tensor*.

These sixteen quantities of the corpuscular theory correspond, in the electromagnetic theory, to the sixteen quantities which are obtained

by multiplying the components of the electric and the magnetic field strengths and which, under a transformation of coordinates, transform in exactly the same way as the products $J_x g_x$, $J_x g_y$, etc. The ensemble of these quantities we denote as the *energy tensor* or as the energy-momentum tensor. The volume density of the electromagnetic energy is determined by the well-known expression

$$\frac{1}{8\pi}(E^2+H^2) = \frac{1}{8\pi}(E_x^2+E_y^2+E_z^2+H_x^2+H_y^2+H_z^2),$$

where $\mathbf{E}$ is the electric and $\mathbf{H}$ the magnetic field strength. In the corpuscular theory the product $J_l g_l = N\epsilon$ corresponds to this 'wave expression'. On going from the original coordinate system 'at rest' to one moving relatively to it, both expressions transform like the square of the time coordinate (t). One substantial difference between them is that while $N\epsilon$ does not vary with the time, E^2 and H^2 oscillate (with a frequency 2ν in the case of homogeneous light of frequency ν) about their average values $\overline{E^2}$ and $\overline{H^2}$, which are equal to half the squares of the amplitudes of $\mathbf{E}$ and $\mathbf{H}$ respectively. The correspondence between the corpuscular and the wave theory of light with regard to its *quantity* or intensity can therefore be expressed in a relativistically invariant (or 'covariant') form by the equation

$$N\epsilon = \frac{1}{8\pi}(\overline{E^2}+\overline{H^2}), \tag{16a}$$

and fifteen further equations of the same kind between the remaining corpuscular quantities $J_x g_x$, $J_x g_y$, etc., on the one hand and the corresponding quantities of the electromagnetic theory of light on the other. We shall here only notice the following three equations:

$$J_x\epsilon = \frac{c}{4\pi}(\overline{E_y H_z}-\overline{E_z H_y}), \qquad J_y\epsilon = \frac{c}{4\pi}(\overline{E_z H_x}-\overline{E_x H_z}),$$

$$J_z\epsilon = \frac{c}{4\pi}(\overline{E_x H_y}-\overline{E_y H_x}), \tag{16b}$$

for the components of the (three-dimensional) energy-current vector (or Poynting vector).

The formula (16) given above is, strictly speaking, not invariant (in the relativity sense), and so can only claim an approximate validity.

II

MATTER

5. Extension of the Wave-Corpuscle Parallelism to Matter

Before Einstein we were content with a purely wave theory of light and a purely corpuscular theory of matter. By his light quantum theory Einstein transferred the corpuscular concept of matter to light, and by combining this concept with the wave concept introduced a quite unexpected dualism or 'parallelism' into the theory of light. Twenty years after this first step was made in combining the theories of light and matter, or of *optics* and *mechanics*, the French physicist, Louis de Broglie, made the second step—the 'complementary' step as it were—in the opposite direction. *He applied the wave conception of light to matter.* In this way there arose in January 1924—for the first time in the history of physics—the idea of 'material waves' which are connected with moving particles of matter.

In his generalization of Einstein's wave-corpuscular parallelism de Broglie was led by purely formal considerations. As we have seen above, this parallelism in Einstein's theory of light, in so far as the *quality* of the light is concerned, is expressed by the relations (15). These relations were set up by Einstein for corpuscles with vanishing rest mass. They remain, however, formally correct, i.e. from the point of view of the relativity invariance condition, also for corpuscles with *non-vanishing rest mass.* Hence de Broglie first attempted to treat light quanta or photons as material particles having a rest mass which is very small but still different from zero. In spite of certain methodological advantages, such a treatment proved to be fundamentally untenable. This attempt, however, led de Broglie quite naturally to the epoch-making idea of applying Einstein's relations (15) to *ordinary material particles*—at first to the elementary particles of matter, i.e. protons and electrons. By means of this small step in a direction in which, until that time, no physicist—apparently not even Einstein himself—had sought to advance, de Broglie came out of the old macroscopic mechanics and discovered the new land which was lying so near and yet had been hidden so long—the land of *microscopic* or *wave mechanics.*

However, de Broglie was unable to penetrate far into this new region. Because of a misleading conception of the connexion between his 'material' or 'phase waves' and the corresponding particles, he came

to a stop almost on the margin. According to his original conception, each individual particle, e.g. a single moving electron, was connected with a system of 'phase waves' in such a way that the path of this particle coincided with a ray of the corresponding wave system. Thus he imagined that the waves, as it were, 'carry along' the particle, and that hence each particle must possess its own train of waves.

The proper method of approach to the newly discovered region of wave mechanics (this expression is due to de Broglie) lay, however, in quite a different direction, and indeed in that direction which is indicated by the relations (16a, b) given above for the wave mechanics of light quanta. These relations, perhaps in a somewhat more general form, must now be applied to material particles and material waves.

In the dual theory of light we find, corresponding to a harmonic train of waves, not just *one* photon, but a *group* of photons distributed in space with a concentration proportional to the square of the amplitude of the waves. In a similar way we must assign to a harmonic system of de Broglie waves, propagated in a definite direction, a homogeneous stream of identical material particles with a density proportional to the square of the amplitude. The two ideas of particles and waves, which at first sight are so foreign to one another, thus become connected by the common idea of *rays*. These rays, which from the corpuscular point of view are to be regarded as the *paths* of the particles and from the point of view of the wave theory as the lines along which the propagation of the waves takes place, are to be so drawn that the number of rays which cross a unit surface perpendicular to them is proportional to the square of the amplitude of the waves at that point.

A system of plane sine de Broglie waves, which are represented throughout the whole of space by formula (14a), corresponds to a homogeneous stream, unlimited in length and cross-section, of identical material particles with the same momentum $\mathbf{g} = h\mathbf{k}$ and the same energy $\epsilon = h\nu$.† The concentration N of these particles must, throughout the whole of space, have the same value—a value proportional to ψ_0^2. The wave function ψ for particles with non-vanishing rest-mass has, in general, a quite different physical meaning than for photons. It has in general nothing to do with the electric and magnetic field strengths which describe the field of light waves.

† It was an important feature of de Broglie's theory that he identified the coefficient h in these equations for *any* material particles with that for photons—that is with Planck's constant—an assumption that has been fully confirmed by experiment. The theoretical implications of this fact and its bearing on the relation between light and matter will be discussed later on.

It should be emphasized that a knowledge of the *mean* number of particles per unit volume (by means of the relation $N \sim \psi_0^2$) in no way allows the position of the separate particles to be determined exactly and their separate motions to be followed. Such an accurate determination of the positions of *discrete* point-like objects by means of a *continuously distributed* quantity (as e.g. the accurate determination of the position of the 'lines of force' which serve to describe graphically a continuous field of force) is obviously fundamentally impossible.

If the length of the de Broglie waves $\lambda = 1/k$ and their vibration frequency ν are known, then we can determine the rest-mass of the corresponding particles as follows:

By eliminating the velocity v from the expressions

$$g = \frac{m_0 v}{\sqrt{(1-v^2/c^2)}} \quad \text{and} \quad \epsilon = \frac{m_0 c^2}{\sqrt{(1-v^2/c^2)}} \tag{17}$$

we obtain the relation

$$g^2 - \epsilon^2/c^2 = -m_0^2 c^2. \tag{17a}$$

From $g = hk$ and $\epsilon = h\nu$ it follows that

$$k^2 - \frac{\nu^2}{c^2} = -\frac{m_0^2 c^2}{h^2}, \tag{18}$$

or

$$m_0 = \frac{h}{c}\sqrt{\left(\frac{\nu^2}{c^2} - \frac{1}{\lambda^2}\right)}.$$

It should be noticed that the quantity

$$w = \nu/k = \nu\lambda \tag{18a}$$

gives the velocity of propagation of the waves. This velocity is given as a function of the wave-length, according to (18), by the formula

$$w = c\sqrt{\left(1 + \frac{m_0^2 c^2}{h^2}\lambda^2\right)}. \tag{18b}$$

In the case of de Broglie waves which are propagated with the velocity of light ($w = c$) we get from (18b) that $m_0 = 0$ which corresponds to the case of light quanta. In this case we can identify the de Broglie waves with the electromagnetic waves.

When $m_0 > 0$, by (18b) the velocity w must be *greater* than the velocity of light. We imagine, for example, a parallel homogeneous bundle of *cathode rays*. From the corpuscular point of view this can be regarded as a stream of electrons which are emitted from the negative electrode of a discharge tube and move with uniform translational velocity. From the wave theory point of view these rays are to be considered as the lines of propagation of certain 'cathode waves' or

'electron waves' in which the electrons play the same role as photons in the case of light waves.

If in (18a) we put $\nu = \epsilon/h$ and $\lambda = h/g$, we get

$$w = \epsilon/g. \tag{19}$$

Hence by (17) we arrive at the remarkable relation

$$w = c^2/v \tag{19a}$$

between the wave velocity and the corpuscular velocity.

Since v, the velocity of the electrons, is always smaller than c, it follows that the velocity of propagation of the cathode waves, as has already been mentioned, must always be greater than the velocity of light—and, indeed, the more so the smaller the value of v. Electrons at rest correspond to waves which travel with infinite velocity, i.e. to vibrations with an infinite wave-length and which accordingly have the *same phase* throughout space. In this case the wave function reduces to the most simple form

$$\psi' = \psi'_0 \cos 2\pi\nu' t', \tag{20}$$

where ν' is the vibration frequency and t' is the time in that system of coordinates in which the electrons are at rest. If we consider the same electrons from another system of coordinates in relation to which they move in the direction of the x-axis with velocity v, then, according to (5), we must, in the above expression, replace t' by

$$\frac{t - xv/c^2}{\sqrt{(1-v^2/c^2)}}.$$

We thus obtain (the law of transformation of the wave amplitude is not taken into account)

$$\psi = \psi_0 \cos 2\pi \frac{\nu'}{\sqrt{(1-v^2/c^2)}}\left(t - \frac{x}{c^2/v}\right), \tag{20a}$$

i.e. the equation of a system of waves with frequency

$$\nu = \frac{\nu'}{\sqrt{(1-v^2/c^2)}}$$

which travel with velocity $w = c^2/v$ in the x-direction. De Broglie first derived formula (19a) in this way. The relation $\nu = \dfrac{\nu'}{\sqrt{(1-v^2/c^2)}}$ corresponds to the relation $\epsilon = \dfrac{\epsilon_0}{\sqrt{(1-v^2/c^2)}}$ $\left(\text{or } m = \dfrac{m_0}{\sqrt{(1-v^2/c^2)}}\right)$ between the energies or masses of an electron in the two coordinate systems ($\nu' = \nu_0$ is the 'rest frequency' of the cathode waves).

The de Broglie relation between corpuscular and wave velocity is not

to be considered as quite new and unexpected. We have indeed seen in § 1 that the refractive index for any rays (the limitation to light rays is not unconditionally necessary) according to the corpuscular theory is v'/v, and according to the wave theory is w/w'. In order that both values may agree—corresponding to the mutual equivalence of both points of view—we must put

$$\frac{v'}{v} = \frac{w}{w'}.$$

If, further, we assume that when $v' = c$ then also $w' = c$, we obtain formula (19a).

It might at first appear that this derivation of formula (19a), which makes no use of the theory of relativity, contradicts the fundamental principle of the theory of relativity—the invariance of the velocity of light. However, this would only be true if, in the case of light rays, we identify the apparent velocity of propagation with the true velocity. With cathode rays we have only one, and indeed the true velocity of propagation, and there is no contradiction to the theory of relativity in assuming different values for this velocity in different regions of space.

6. Interference and Diffraction of Cathode Rays

Fizeau's experiment was not the only nor even the most important argument against Newton's corpuscular theory of light. Interference phenomena ('Newton's rings') first discovered by Newton himself and especially the diffraction phenomena discovered and investigated by Young and Fresnel decided the complete (and apparently final) victory of the wave theory. For from the point of view of the wave theory all these phenomena could be explained in a quite straightforward way, even down to the finest quantitative details, while from the point of view of the old corpuscular theory of light they remained completely incomprehensible.

The introduction of relativity mechanics instead of Newton's mechanics and the corresponding remodelling of Newton's corpuscular theory of light by Einstein altered this state of affairs only a little. The relativity conservation laws of energy and momentum in conjunction with the wave-corpuscular relations $g = h/\lambda$, $\epsilon = h\nu$ for the light rays only explained phenomena connected with the *quality* of the light (photo-electric effect, Compton effect). To interpret phenomena depending on the *quantity* of light—for instance, interference and diffraction—and to calculate the corresponding measure of quantity, e.g. the

intensity of the reflected, transmitted, and diffracted rays, we must proceed from the wave theory laws of the propagation of light, and in order to translate these results into corpuscular speech must use the wave-corpuscular relation $N \sim \psi_0^2$. We could say, therefore, that while the phenomena of interference and diffraction of light did not contradict the corpuscular conception, yet they were only explainable quantitatively in terms of the wave conception.

The transference of the wave-corpuscle parallelism from light rays to material rays undertaken in § 5 would be unjustified if we had to limit ourselves to phenomena dependent on 'quality' which can already be explained purely corpuscularly. To justify this far-reaching generalization we must see if we can find for rays of matter interference and diffraction phenomena, i.e. phenomena dependent on quantity which can only be explained by a wave theory.

Such experiments have, in fact, been carried out recently with cathode rays, and indeed with positive results for both groups of the wave-corpuscule relations—i.e. the de Broglie 'quality relation' $g = h/\lambda$ and the quantity relation $N \sim \psi_0^2$.

It should be noticed that the possibility of an experimental observation of interference of light rays depends on their wave-length, or more exactly on the ratio of their wave-length to the distances between the elementary parts of our optical apparatus. In a diffraction grating, for example, to observe interference conveniently, the distance between the elementary 'diffracting regions' must be of the same order of magnitude as the wave-length of the light under investigation.

Thus before taking up interference experiments with material waves it is necessary to have some idea of the order of magnitude of the corresponding wave-lengths.

If in the de Broglie formula $\lambda = h/mv$ we put values of some such magnitude as $m = 1$ gm. and $v = 1$ cm. sec.$^{-1}$, we obtain

$$\lambda \cong h \cong 6 \times 10^{-27} \text{ cm.}$$

This wave-length lies outside of the region accessible to observation. To make λ as large as possible we must take particles with the smallest possible mass and the smallest possible velocity. Thus we have the most favourable conditions with slow cathode rays. Here $m = 9 \times 10^{-28}$ and hence $\lambda \cong 7{\cdot}3/v$ cm. The slowest cathode rays with which exact experiments can be carried out are produced by a potential difference V of the order of magnitude of a volt. According to the formula $\frac{1}{2}mv^2 = eV/300$ this corresponds to a velocity $v = 10^8$ cm. sec.$^{-1}$ and to

a wave-length $\lambda \cong 10^{-7}$ cm. We have in general (for an accelerating potential difference which is not too high)

$$\lambda = \frac{10^{-7}}{\sqrt{V}} \text{ cm.}$$

With the usual potentials from 1 to about 10,000 volts, cathode rays have a wave-length from 10^{-7} to 10^{-9} cm., i.e. approximately the same wave-length as Röntgen rays. To observe interference of Röntgen rays we use crystals which, for this purpose, can be regarded as natural three-dimensional diffraction gratings. Hence it is to be expected that the same crystals can also be used as diffraction gratings for cathode rays and that we shall find interference phenomena similar to those displayed by Röntgen rays. *Metal crystals* have proved specially suitable for interference experiments with cathode rays.

We shall here briefly discuss the most important experiments in this new field.

1. *The experiment of Davisson and Germer* (1927). A bundle of homogeneous cathode rays with a de Broglie wave-length of the order of magnitude of 10^{-8} cm. was investigated by the Bragg method of reflection. A series of reflection maxima were observed and indeed with *almost* exactly the same angles of reflection (and incidence) as in the case of Röntgen rays of the same wave-length. The small deviations from the Bragg formula

$$2d \sin\theta = n\lambda \tag{21}$$

(d = distance between the lattice planes, θ = glancing angle between these planes and the incident or reflected ray, $n = 1, 2, 3, \ldots$) can be explained very satisfactorily by the refraction effects appearing at the surface. With Röntgen rays the refractive index μ is practically equal to 1. With cathode rays, on the other hand, it can be considerably different from 1. In the Bragg formula (21) we must then replace the angle θ by the corresponding angle θ' for the refracted rays and λ by $\lambda' = \lambda/\mu$. Now

$$\mu = \frac{\sin(\frac{1}{2}\pi - \theta)}{\sin(\frac{1}{2}\pi - \theta')} = \frac{\cos\theta}{\cos\theta'},$$

i.e.
$$\sin\theta' = \sqrt{(1-\cos^2\theta')} = \sqrt{\left(1 - \frac{1}{\mu^2}\cos^2\theta\right)}.$$

The corrected Bragg formula

$$2d\sqrt{(\mu^2 - \cos^2\theta)} = n\lambda \tag{21a}$$

completely agrees with the experimental results of Davisson and Germer. For the refractive index μ which corresponds to the transition

of cathode rays from vacuum into the various metal crystals (a crystal of nickel was first used) one obtains values which are greater than 1 and which, with increase of velocity, approach this as a limiting value. This means that electrons, on entering the metal, experience an *acceleration*. Hence, on escaping from a metal, they must be slowed down. This result is in general agreement with the fact that the so-called 'free electrons' (which are responsible for the electrical conductivity of metals) ordinarily remain bound in the metal. Only with high temperatures do they obtain a sufficiently large kinetic energy to overcome the surface forces.

According to the corpuscular theory the refractive index for cathode rays which are not moving too fast is given by the formula

$$\mu = \sqrt{\left(1 + \frac{U}{\frac{1}{2}mv^2}\right)}. \qquad \text{[See (2), § 1.]}$$

Here U is the work done by the force mentioned on entry into the metal. We can express this in volts by the relation $U = \frac{eV_0}{300}$. Thus with $\frac{1}{2}mv^2 = \frac{eV}{300}$ we get

$$\mu = \sqrt{\left(1 + \frac{V_0}{V}\right)}. \tag{21b}$$

This formula agrees with the results of Davisson and Germer and enables us to determine the internal potential V_0 of the various metals. In this way values between 13 and 20 volts are obtained.

2. *The experiments of G. P. Thomson* (1928). A bundle of fast cathode rays was sent through a very thin metal foil (Al, Au, etc.). Such a foil consists of a crowd of crystals oriented at random. With Röntgen rays we obtain the well-known Debye-Scherrer diffraction picture which consists of a number of concentric rings.

Thomson obtained the same picture with cathode rays. The wave-length determined from the diameters of the interference rings agreed exactly with the de Broglie wave-length $\lambda = h/mv$. Similar experiments with slow cathode rays were carried out somewhat later by Rupp with a series of foils of various metals (Al, Cu, Au, Pb, Ni, etc.). The results were the same as in the experiments of Thomson.

3. Recently it has been possible to observe interference of Röntgen rays by an ordinary optical diffraction grating. Because of the large distance d between neighbouring lines of such a grating compared with the wave-length of the Röntgen rays, interference phenomena can only

be observed with very oblique incidence, i.e. with very small angles θ. Rupp carried out similar experiments in the summer of 1928 with slow cathode rays and obtained exactly the same results as with Röntgen rays. The reality of cathode waves became finally verified by this experiment. Measurement of the wave-length gave values which agreed exactly with the theoretical formula $\lambda = h/mv$.

4. The fact that not only cathode rays but also atomic and molecular beams are associated with waves, has been demonstrated experimentally by *Dempster* for hydrogen atoms, and by *Stern* for heavier atoms by the method of crystal reflection, the crystal acting in this case not as a volume but as a surface (two-dimensional) diffraction grating. The de Broglie relation between velocity and wave-length has been verified in all cases.

The discovery of the wave nature of cathode and other material rays is one of the most impressive examples of the power of the human intellect and of the wonderful force and fruitfulness of the theory of relativity.

7. Group Velocity of Wave Processes and the Theory of Wave Packets

The earlier struggle between the corpuscular and the wave theory of light was decided by Fizeau's experiment in favour of the latter. Because it is concerned, not with the *true*, but with the *apparent* velocity of light, Fizeau's result could not serve as an argument against the new corpuscular theory of light of Einstein. Thus for this theory there arose the task of following in detail the true motion of light quanta in a material body taking into account their absorption and re-emission.

With rays of matter the facts are somewhat different. Here there is no question of an apparent wave velocity. It would therefore seem possible, by direct measurement of the velocity of cathode rays by Fizeau's method and by comparing the ratio of velocities with the refractive index, to decide the question as to whether the cathode rays are to be regarded as wave rays or as corpuscular rays.

We must, however, remember that, even with the assumption of the wave nature of the rays, the velocity directly determined by Fizeau's method must, in general, be different from the true 'phase velocity' w of the waves. This circumstance, which was first discovered by Lord Rayleigh in investigating water waves, is due on the one hand to the *inhomogeneity* of the rays which in reality is always present, and on the other hand to their *dispersion*, i.e. to the dependence of the 'phase

velocity' of completely homogeneous rays on the vibration frequency (or the wave-length).

Rays which are regarded as practically 'homogeneous' are in fact not so, but must be treated as a 'group' of strictly homogeneous rays whose frequencies ν and wave numbers k are continuously distributed throughout very small regions $\Delta\nu$ and Δk respectively. Such a ray or wave group can be represented analytically by the integral

$$\psi(x,t) = \int_{\nu_0-\frac{1}{2}\Delta\nu}^{\nu_0+\frac{1}{2}\Delta\nu} a_\nu e^{i2\pi(kx-\nu t)}\, d\nu, \tag{22}$$

where a_ν is a slowly varying function of ν in the interval $\Delta\nu$. k is to be regarded as a definite function of ν.

We shall assume that the interval $\Delta\nu$ is so small that, for the purposes of integration, we can put $a_\nu = \text{const.}$ and

$$k = k_0 + \left(\frac{dk}{d\nu}\right)_0 (\nu-\nu_0).$$

We then get with $\nu-\nu_0 = \xi$

$$\psi = a_\nu e^{i2\pi(k_0 x-\nu_0 t)} \int_{-\frac{1}{2}\Delta\nu}^{+\frac{1}{2}\Delta\nu} e^{i2\pi\left[\left(\frac{dk}{d\nu}\right)_0 x-t\right]\xi}\, d\xi,$$

or

$$\psi = \frac{a_\nu}{\pi} \frac{\sin \pi\Delta\nu\left[\left(\frac{dk}{d\nu}\right)_0 x-t\right]}{\left(\frac{dk}{d\nu}\right)_0 x-t} e^{i2\pi(k_0 x-\nu_0 t)}. \tag{22a}$$

We can regard the vibration process represented by this formula as a homogeneous train of waves (frequency ν_0, wave number k_0) with a *variable amplitude* which possesses a strong maximum ($= a_\nu \Delta\nu$) when $\left(\frac{dk}{d\nu}\right)_0 x-t = 0$, i.e. when $x = \left(\frac{d\nu}{dk}\right)_0 t$. The maximum amplitude and, in general, all points with a given constant amplitude travel, therefore, with a 'group velocity' different from the phase velocity—a velocity which, on neglecting the suffix 0, is given by the formula

$$w_g = \frac{d\nu}{dk}. \tag{23}$$

This result can be seen immediately if we notice the fact that in the propagation of two vibrations somewhat 'out of tune', not only time but also spatial beats must arise. Just as the number of the former in unit time is $d\nu$, so the number of the latter in unit length is dk. The

velocity of propagation of these space-time beats—i.e. the group velocity—is obviously equal to the ratio $\frac{d\nu}{dk}$.

This, and not the phase velocity $w = \nu/k$, is observed in Fizeau's experiment.

If we put $\nu = wk$ in (23) we obtain

$$w_g = w + k\frac{dw}{dk} = w - \lambda\frac{dw}{d\lambda}. \tag{23a}$$

The differential coefficient $\frac{dw}{dk}$ $\left(\text{or } \frac{dw}{d\lambda}\right)$ determines the *dispersion* of the waves in the given region. If dispersion is absent the group velocity coincides with the phase velocity. Otherwise, however, they are different.

For light rays this difference is in general very small, so that we can practically neglect it. (We are here clearly concerned with the macroscopic theory of the propagation of light in material bodies and with the apparent phase and group velocities respectively.)

It is quite otherwise with de Broglie's material waves. The dependence of the (true) phase velocity on the wave-length is here expressed by formula (18b). Thus we have by (18)

$$k\,dk = \frac{\nu\,d\nu}{c^2},$$

i.e.

$$\frac{d\nu}{dk} = c^2\frac{k}{\nu} = \frac{c^2}{w},$$

or by (19a)

$$w_g = v. \tag{24}$$

The group velocity of material waves coincides, therefore, with the velocity of the corresponding particles.

This result can also be obtained as follows: In formula (23) we replace ν and k by the quantities ϵ and g proportional to them and consider their differentials as infinitesimal alterations during the time dt. We then get

$$d\epsilon = \frac{dg}{dt}w_g\,dt.$$

Now dg/dt—the change of the momentum in unit time—must be equal to the force f acting on the particle. On the other hand the alteration of the energy $d\epsilon$ must be equal to the work done by this force, i.e. to the product $fv\,dt$. Hence it follows that $w_g = v$.

We see, therefore, that with material rays, e.g. cathode rays, de Broglie's wave theory, with due allowance for the unavoidable hetero-

geneity of these rays, leads to the same observable velocity of propagation as the usual corpuscular theory. Fizeau's experiment, applied to such rays, is therefore useless for deciding between the two theories. This experiment has in reality not been carried out. Yet we can maintain with considerable certainty that its result would be in agreement with the corpuscular theory. A contradiction to the wave theory could only arise on the naïve assumption that the observed velocity is to be identified with the 'phase velocity'. If, on the other hand, we introduce the idea of group velocity, we at once remove this contradiction and both theories appear to be 'correct'.

The fact that for light Fizeau's experiment could serve as a crucial experiment is explained by the relatively small dispersion of light waves and the resulting indistinguishability of the phase and group velocities. It must, however, be remembered that we are here concerned with 'apparent' macroscopic velocities. A consistent application of Einstein's theory of light quanta to the propagation of light in material bodies must—as mentioned above—lead to the same results as the application of the electromagnetic theory of light.

We have seen that the group velocity of waves can be taken to be the velocity of propagation of local maxima of amplitude, these maxima being produced by superposition of trains of harmonic waves of various wave-lengths. In measuring the group velocity of cathode waves we thus measure the velocity of these local maxima and obtain the velocity attributed to the separate electrons. It seems natural therefore to associate the electrons with such local maxima.

This association induced de Broglie to believe that an individual particle should be considered not as a special 'addition' to the waves which themselves were essentially different from it, but as a local sharp maximum of amplitude in a wave system. Such a formation *Schrödinger* called a 'wave packet'.

A wave packet can be constructed by superimposing a group of spatially unlimited plane harmonic waves whose frequencies, wave-lengths, and *directions of propagation* vary continuously over very narrow ranges. A wave packet is thus nothing else than a 'wave group' (in the sense of the definition of this section) which, however, is limited not only in *one* direction but on *all* sides. This explains why such a packet moves in space with the group velocity.

The amplitude of vibration in a wave packet reaches a maximum value at the middle point. On all sides it falls away gradually, and indeed so much the faster the larger the values of the spectral interval

and the range of angles of the constituent elementary waves. De Broglie showed that it is possible in this way by a limiting process to construct wave packets with the amplitude becoming infinite at a definite point moving with the group velocity. Such singular points represent, on his theory, material particles devoid of spatial extension, for instance, electrons considered as point-charges. To interpret the spatial extension ('radius') usually attributed to electrons (on the basis of the electromagnetic theory of mass) Schrödinger, to whom the further development of de Broglie's ideas is due, preferred, however, to represent electrons by wave packets of finite dimensions.

The latter point of view is, however, contradicted by the fact that wave packets in general do not preserve the same dimensions, but spread out in space as time goes on, whereas the dimensions of the particles they represent remain constant (see below).

But even apart from this, the conception of particles as wave packets contradicts the facts of observation on the interference and diffraction of the corresponding rays. If we imagine a bundle of cathode rays to be a collection of wave packets, each wave packet corresponding to a separate electron, then the scattering of these rays at a diffraction grating would not produce the usual interference pattern similar to that which is observed in the case of homogeneous light waves. In fact to explain this interference pattern by the wave theory we must assume that the incident rays form a sinusoidal train of waves with a *constant amplitude* throughout space (disregarding of course the finite cross-section and divergence of the bundle of rays). The action of a diffraction grating is as follows: By the superposition (interference) of the elementary waves of the *same* amplitude which are emitted by the separate lines (or planes) of the grating resultant vibrations are produced with a spatially constant amplitude depending only upon the phase difference, that is, upon the direction, of the reflected beams.

Thus the representation of the individual electrons in a beam of cathode rays by wave packets is quite unable to explain diffraction and interference phenomena. The influence of the grating would manifest itself in this case in the reflection and eventually in the spreading of the separate wave packets, more or less in accordance with the corpuscular theory, and no trace of the actually observed interference and diffraction phenomena would be displayed.

Discontinuous elementary actions of light or material rays (such as the photo-electric effect), which are interpreted as evidence for their corpuscular structure, cannot, therefore, be explained by a corpuscular

structure of the corresponding wave field. They must also occur with a completely homogeneous structure—or rather a complete absence of structure—of this field.

We thus see that the *reduction* of particles to waves made so tempting by the coincidence of particle velocity with group (or wave packet) velocity *is impossible*, and that this coincidence must have an entirely different physical meaning.

8. The Probability Concept

In dealing with the two aspects (wave and corpuscular) of the rays of light we did not try to reduce one of them to the other but considered them as *two parallel series* which are related to one another by a *purely symbolic* correspondence—a correspondence similar to that existing between the physiological processes in our brain and the accompanying psychological processes. This symbolic correspondence was expressed mathematically by the 'quality relations' $\epsilon = h\nu$, $\mathfrak{g} = h\mathfrak{k}$ and the 'quantity relation' $N \sim \psi_0^2$. The latter relation is convenient so long as we are dealing with very intense rays, i.e. rays corresponding to very large values of N (number of particles per unit volume). It becomes, however, meaningless for very weak rays, e.g. when N is a small fraction of unity. We are then confronted with the question of the location of the separate particles in a much more pressing manner than when N is large. Now it is clear that this question cannot be answered in a definite way. For the association of *discontinuous* (point-like) corpuscles with a *continuous* wave field by means of the relation $N \sim \psi_0^2$ involves an *indeterminacy* with respect to the positions of the corpuscles. (A similar situation arises in the graphical representation of a field of force by lines of force.) If ψ_0^2 is known throughout space (as a continuous function of the coordinates), then we can only calculate the *average* or *probable* number of corpuscles present in a given volume ΔV. The product $\psi_0^2 \Delta V$ which determines this *probable number of similar particles* can thus be considered as a *measure of the probability* of the presence of *one of these particles* in the volume element ΔV, and hence the quantity ψ^2 can be defined as a measure of the *probability density* for the presence of a *single* particle at the point under consideration. If the function ψ can be normalized in such a way that the integral $\int \psi_0^2 \, dV$, taken over the whole space, is equal to 1, then $\psi_0^2 \, dV$ will be the absolute value of the probability of finding a particle in the volume element dV; in the contrary case (for instance, in the case of plane harmonic waves extending throughout all space) ψ_0^2 as a function of the coordinates will

determine the *relative* probability for finding the particle at different points.

This modified form of the 'quantity relation', which can be denoted as the *probability relation*, was introduced by *Born*. It has an important advantage over the original form, inasmuch as it can be applied to a *single particle*. Our considerations are thus no longer limited to material beams consisting of a number of particles of the same kind moving in the same way, and exerting no action on one another. *So long as this mutual action is excluded or can be neglected,* that is, so long as the different particles are moving independently, it appears unnecessary to consider them as one whole.† Picking out one of them we can follow its career without paying attention to the rest. The effect produced by all of them will be simply the sum of the elementary effects produced by the individual particles.

The question now arises as to how these elementary effects are determined. This question obviously reduces to that of determining the motion of a single particle under given conditions (e.g. in a given field of force), that is, determining its position at any time. In the old 'classical' mechanics this question was answered by obtaining the coordinates of the particle x, y, z, as certain functions of the time, these functions being determined by integrating Newton's equations of motion, taking into account the initial conditions (as to position and velocity). In the new mechanics, however, in which the conception of a particle is associated with the conception of a system of waves through the idea of probability, an answer of this kind to the above question is impossible. The *deterministic* description of the motion of a particle, implying the possibility of its exact location at any time, must here be abandoned. We must confine ourselves to the less ambitious task of determining the probability of finding the particle in any place at a given time. Since this probability is measured by the square of the amplitude of the waves associated with the particle, we must determine the laws of propagation of these waves, or more exactly the form of the 'wave function' $\psi = \psi(x, y, z, t)$. The classical mechanical problem which consisted in finding x, y, z as functions of t is thus replaced by a wave problem in which x, y, z, and t must be treated as independent variables.

It is clear that under such circumstances the determination of the motion of a particle in the usual sense of the word, that is, in the sense of tracing the change of its position with time, becomes impossible.

† We shall see later on that this is not exactly true (Ch. IV).

The deterministic description of the motion of a particle—and consequently of all the effects connected with it—must be given up and replaced by a probability description based on the determination of the corresponding wave system.

This is the fundamental idea of the new wave mechanics which can therefore be defined as the *probability theory of the motion of material particles*. We shall illustrate it by some examples relating to reflection, transmission, and diffraction of light or cathode rays. These effects were formerly described either in terms of waves or in terms of a *stream* of similar particles. We can now consider them as a sum or 'superposition' of elementary effects due to the separate particles, associating each of the latter with the same type of wave system. The wave systems referring to different particles may differ from one another by the phase only. If these phases, or rather phase differences, are distributed in an irregular manner (incoherent waves), then the square of the amplitude of the resulting vibration will be equal to the sum of the squares of the amplitudes of the various elementary wave systems—corresponding to the simple addition of the number of particles associated with them.

Suppose an elementary system of waves falls on a plane surface where it is partially reflected and partially transmitted. We can describe this effect from the corpuscular point of view by saying that the particle associated with the incident wave on impinging on the surface has a certain probability of being reflected, that is, of becoming associated with the reflected wave, and a certain probability of being transmitted, that is, of becoming associated with the transmitted wave, the ratio of these probabilities being equal or proportional to the ratio of the squares of the amplitudes of the reflected and transmitted wave systems. This description is, in principle, identical with that given by Newton for the case of light; the only difference is that instead of speaking of *probabilities* Newton spoke of 'fits' of reflection and transmission, and so attributed to the light particles a property similar to 'free will'.

It must be emphasized that the above description of the phenomena of reflection and transmission both of light and of cathode (or any other) rays is no mere speculation, but is an adequate interpretation of the experimental facts. If we could project at the surface similar particles (photons, electrons, etc.) at intervals of time sufficiently long for each particle to be reflected or transmitted before the arrival of the next one, we should observe that under apparently identical conditions

some particles would be reflected and others transmitted. Their relative numbers could then be regarded as the measure of the probability of the corresponding issue for each individual particle. This probability would agree with that calculated from the amplitude of the associated waves.

Such an experiment is not beyond the range of present-day experimental physics which can readily register the elementary effect due to a single electron or photon (by means of Geiger's counter, for instance). In Newton's time an 'elementary experiment' of this kind was of course impossible. His theory of light particles with their odd 'fits' of reflection and refraction was therefore arbitrary and artificial, for all the optical phenomena then known could have been simply and naturally described and explained by the wave theory of light without introducing the idea of any particles associated with these waves. Since particles of ordinary matter seemed to behave sanely, in agreement with the mechanical laws set forth by Newton himself, and without 'fits' of any kind, it was natural to deny the existence of particles which would display such fits.

The present position, however, is entirely different, for we now know that this view was wrong and that electrons do display the same sort of 'fits' as those postulated by Newton for his light particles.

It is true that an experiment of the type described above for observing the reflection or transmission of a single electron has not yet been performed. But experiments on the diffraction of cathode rays point to the same conclusion unless we are prepared to give up the corpuscular conception of matter as it was given up before for light.† Here, as before, the relative number of electrons diffracted by a grating or a crystal in a given direction is a measure of the probability of such a diffraction for a single electron (since the electrons move independently) and is determined by the amplitude of the corresponding diffracted waves. If we could carry out diffraction experiments with cathode rays so weak that the elementary effects, for instance the incidence of the separate electrons on a fluorescent screen, would be separated from one another by sufficiently long intervals of time, then we should undoubtedly observe the different electrons to strike some places of the screen often and other places more rarely. There would be an apparent lack of regularity—a lack of determinateness—for the individual acts, which, however, would produce in the long run a regular

† Actually Newton was the first to discover interference phenomena ('Newton's rings'), which he did not hesitate to interpret in terms of his corpuscular theory with the help of his idea of 'fits'.

distribution of intensity, determined by the amplitude of the diffracted waves at the corresponding places.

The above examples show that the introduction of the idea of probability into elementary physical processes by means of Born's probability relation has far-reaching implications, extending far beyond this relation. In considering, for instance, the phenomena of reflection and refraction from the 'corpuscular-statistical' point of view, we come upon the problem of determining the probability of *transition* from one type of motion, corresponding to the incident particle, to another type of motion, corresponding to the reflected or transmitted particle. The determination of these *transition probabilities* reduces to the determination of the 'location probabilities' by means of Born's relation, and of course to the determination of the laws of reflection and transmission of the waves associated with the particles under consideration.

The indispensability of the idea of 'transition probabilities' for the description of elementary physical processes was first recognized by Einstein—who in this and many other respects may be regarded as Newton's successor. Indeed the idea of transition probabilities corresponds exactly to Newton's 'fits'. In his famous work on the mutual action of matter and radiation (1917) Einstein introduced, for the first time, the idea of transition probabilities for absorption and emission of light. Only by wave mechanics has it become possible actually to calculate these transition probabilities.

As stated above, wave mechanics consists in describing the behaviour of a particle by means of probabilities derived from the determination of the associated waves. Thus the idea of probability provides the connecting link between the parallel ideas of particles and waves. If we wish—as Born and many others do—to stress the corpuscular idea, i.e. to interpret physical phenomena in terms of the motion of discrete corpuscles, then we can regard the corresponding waves merely as 'probability waves' which possess no immediate physical reality and have only an auxiliary meaning. To me, personally, the opposite point of view in which we regard the waves as the fundamental, primary phenomena and the corpuscles as secondary seems more satisfactory. It is certainly not correct to consider light waves, that is, an electromagnetic field, merely as a probability field for photons. One has only to think of the simplest special case of an electromagnetic field, i.e. of the electrostatic field, to see that the introduction of corresponding corpuscles is in general impossible. Because of the deep analogy between light rays and material rays, it seems to be necessary to ascribe

to de Broglie's waves associated with material particles an immediate physical reality as well.

Be that as it may, we must reject the principle of determinism or causality with respect to the particles *but preserve it for the laws of propagation of the associated waves*. Otherwise an exact determination of the wave function (or functions) would be fundamentally impossible. The introduction of the idea of probability occurs, therefore, only in the corpuscular interpretation of the wave processes and thus has somewhat the same significance as the introduction of 'free-will' in the psychological interpretation of physiological processes in the brain (which are assumed to be primary and completely determined). Independently of whether we regard material waves merely as probability waves or ascribe a physical reality to them, we must treat physical occurrences, considered from the corpuscular point of view, as indeterminate. The earlier deterministic mechanics only recognized the probabilities 1 and 0. Its purpose was to determine those events whose probability is equal to 1, and it did this by integrating certain equations of motion, taking into consideration exactly known initial conditions. The new mechanics does not recognize exact initial conditions. The classical alternative—either necessary or impossible—it does not recognize at all. For the new mechanics *all events* are to be regarded as *possible*. For it there is no such thing as events with the probability 1 or 0. Only intermediate probabilities—greater than 0 and less than 1—come into consideration. Hence the new mechanics is concerned, not with the determination of certain unique events which shall have a given probability (= 1), but with the determination of the probability of all imaginable (i.e. possible) events. Thus if the probability function reaches a sharp maximum the corresponding event (as well as an event differing only a little from it) is extremely probable—although *not* completely necessary—and is usually observed.

In those macroscopic processes which we usually observe, and in which an enormous number of elementary systems are concerned, this 'elementary indeterminacy' scarcely manifests itself at all. The 'law of large numbers' prevails here—a law which, just as for human society, completely effaces the accidental, i.e. the individual and uncontrollable, and probability statements take on a statistical conformity to law.

Because of the apparently determinate character of macroscopic (i.e. statistical) processes, it was believed that the corresponding elementary processes were also strictly determinate. However—as Exner (1919) had hinted even before the coming of wave mechanics—this is

a completely unfounded prejudice from which we are only now beginning to emancipate ourselves.

9. Wave Packets as Probability Packets; Uncertainty Relations

The introduction of the idea of probability to link together the wave and the corpuscular conception throws a new light upon the coincidence between the group velocity of material waves and the velocity of the corresponding particles. As stated in § 7, this coincidence led de Broglie (who discovered it) to regard a particle *as* a wave packet. We can now say, however, that a wave packet does not *constitute* a particle but does *represent* a particle confined to a more or less limited region of space, in the sense that the probability of finding the particle in the region occupied by the packet is large and outside this region very small (or zero). From this probability point of view it is perfectly natural that the packet, or rather its centre, should move with the same velocity as that attributed to the particle.

The theory of wave packets acquires in this connexion a very important physical meaning, for it enables us to make a closer comparison of classical deterministic mechanics with wave or probability mechanics.

So long as we consider plane harmonic waves extending throughout the whole of space (or one-half of it for reflection and refraction), such a comparison is difficult because the position of the particle associated with the waves is completely indeterminate, all positions being equally probable. It must be emphasized, however, that the *velocity* of the particle as determined by the length and the direction of propagation of the corresponding waves has, in this case, a perfectly definite value.

If, however, we consider a particle associated with a wave packet, the indeterminateness of position is limited (at any given instant of time) to a more or less small region of space, and this makes it easier to compare its motion with that defined by classical mechanics. *This increase in the accuracy of locating the particle is achieved, however, at the cost of introducing an indeterminateness in the value and direction of its velocity.* In fact, as we have already seen, a wave packet can be constructed by superposing a number of plane harmonic wave systems with wave numbers and directions of propagation varying within certain small limits. So far then as we can judge of the velocity $\mathbf{v}$ of the particle associated with the packet under consideration, by applying the relation $m\mathbf{v} = \mathbf{g} = h\mathbf{k}$ to the constituent plane harmonic wave systems, we have to admit that the momentum $\mathbf{g} = m\mathbf{v}$ of the particle and consequently its velocity remains to a certain extent indeterminate.

It can further easily be shown that this indeterminateness increases when the linear extension of the wave packet decreases, that is, when the indeterminateness or 'uncertainty' in the *position* of the particle in space becomes smaller.

For the sake of simplicity we shall consider a 'linear' wave packet, that is, a wave group formed by superposing harmonic wave systems with the same direction of propagation and with constant amplitudes, but with different frequencies and wave numbers, continuously varying within the intervals Δk, $\Delta\nu$. This group of waves propagates itself like a harmonic train of waves with the mean frequency ν_0 and wave number k_0, and a variable amplitude

$$A \sim \frac{\sin \pi\Delta\nu\left[\left(\frac{dk}{d\nu_0}\right)x-t\right]}{\left[\left(\frac{dk}{d\nu}\right)_0 x-t\right]} = \frac{\sin \pi(\Delta k\,.\,x-\Delta\nu\,.\,t)}{\left[\left(\frac{dk}{d\nu}\right)_0 x-t\right]}$$

(see equation (22a)). This, of course, is not the general form of A as a function of x and t, but the simplest form, corresponding to the assumption that the interval $\Delta\nu$ (or Δk) is very small and that a_ν is constant within this interval and vanishes outside. Such a wave group may be referred to as *elementary*. In the general case $\Delta\nu$ need not be small, and a_ν can be an arbitrary function of ν, having large values only in the neighbourhood of ν_0.

For a constant value of t, for instance $t = 0$, A^2 is a function of x which reaches its principal maximum when $x = 0$. Beyond the first two zero points at $\Delta k\,.\,x = \pm 1$, A^2 has a series of secondary maxima which, however, can be neglected in comparison with the principal maximum. The *effective length* Δx of the resulting train of waves, i.e. the length of that part in which A^2 has values considerably different from zero, is therefore given by the relation

$$\Delta x\,.\,\Delta k \cong 1. \tag{25}$$

The corpuscular meaning of this relation is obtained if we regard A^2 as a measure of the probability of the presence of the corresponding particle at the point concerned and $hk = g$ as its momentum. We thus get

$$\Delta x\,.\,\Delta g \cong h, \tag{25a}$$

or, if the mass m of the particle can be regarded as constant (that is, if its variation with velocity v may be neglected),

$$\Delta x\,.\,\Delta v \cong \frac{h}{m}. \tag{25b}$$

It must be remarked that Δx denotes here not the actual width of the 'uncertainty region'—*which theoretically is always infinite*—but only its *effective* width, the probability of finding the particle outside this effective region being relatively small (but not exactly zero).

A wave packet can be formed by superposing harmonic plane wave systems with values of all three components k_x, k_y, k_z of the wave number vector $\mathbf{k}$ varying continuously within small intervals Δk_x, Δk_y, Δk_z [k being determined by the relation $k = \sqrt{(k_x^2+k_y^2+k_z^2)}$]. It can easily be seen that the wave function ψ representing such a packet is equal to the product of the three functions ψ_x, ψ_y, and ψ_z representing the three wave groups in the directions of the respective coordinate axes. We thus obtain in this case the following generalizations of the preceding relations

$$\Delta x.\Delta k_x \cong 1; \qquad \Delta y.\Delta k_y \cong 1; \qquad \Delta z.\Delta k_z \cong 1 \tag{26}$$

$$\Delta x.\Delta g_x \cong h; \qquad \Delta y.\Delta g_y \cong h; \qquad \Delta z.\Delta g_z \cong h \tag{26a}$$

$$\Delta x.\Delta v_x \cong \frac{h}{m}; \qquad \Delta y.\Delta v_y \cong \frac{h}{m}; \qquad \Delta z.\Delta v_z \cong \frac{h}{m}. \tag{26b}$$

If we consider the above expression for the amplitude A of a wave group (or the corresponding expression for the effective amplitude of a wave packet) as a function of the time, keeping the value of x (and of both the other coordinates) fixed, then we also obtain a relation

$$\Delta t.\Delta \nu \cong 1 \tag{27}$$

which connects the 'duration' of the group (or packet) with its spectral breadth $\Delta\nu$ measured in units of frequency. If instead of the frequency ν we introduce the energy $\epsilon = h\nu$ of the corresponding particle, we get Heisenberg's fourth uncertainty relation

$$\Delta\epsilon.\Delta t \cong h. \tag{27a}$$

After what has been said previously, we need not discuss this relation in detail.

In deriving the relations (26a) and (27a) we commenced with the given spectral widths Δk or $\Delta\nu$ and estimated the effective space or time breadths of the wave group (or wave packet). The complete symmetry of these relations with regard to the corresponding pairs of quantities suggests that they can also be derived in the reverse direction, i.e. if we begin with a wave group (or packet) with given space and time breadths and estimate the corresponding effective spectral breadths in wave numbers and frequencies.

As a matter of fact, if we consider a train of waves travelling in the

direction of the x-axis, then we can only treat it as a harmonic system of waves if its amplitude has the same value throughout all space. If, on the other hand, this amplitude is only different from zero within the finite region Δx, we have, strictly speaking, no harmonic train of waves in itself, however large Δx may be compared with the wavelength λ_0. By Fourier's integral theorem, however, we can represent such a 'limited train of waves' as the superposition of infinite and therefore strictly harmonic trains of waves with all possible wave numbers from $k = -\infty$ to $k = \infty$; they will, of course, have different amplitudes, with the maximum amplitude corresponding to $k = k_0 = 1/\lambda_0$.

For any function $f(x)$ which is continuous or has only finite discontinuities and behaves suitably at infinity we have

$$f(x) = \int_{-\infty}^{+\infty} A_k e^{2\pi i k x}\, dk,$$

where the amplitude coefficient A_k is given as a function of k by the formula

$$A_k = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i k x}\, dx.$$

If we apply this formula to a function of x defined by

$$\begin{aligned} f(x) &= e^{2\pi i k_0 x} \qquad (|x| < \tfrac{1}{2}\Delta x), \\ &= 0 \qquad\qquad (|x| > \tfrac{1}{2}\Delta x), \end{aligned}$$

we obtain

$$A_k = \int_{-\frac{1}{2}\Delta x}^{+\frac{1}{2}\Delta x} e^{2\pi i (k_0 - k)x}\, dx = \frac{\sin \pi (k_0 - k)\Delta x}{\pi (k_0 - k)}.$$

This expression has its principal maximum (Δx) at $k = k_0$. When $(k_0 - k)\Delta x = \pm 1$ it becomes equal to zero and can be neglected outside the interval $|k - k_0| = 1/\Delta x$. The effective spectral breadth of the limited wave train is thus connected with its length Δx by the same relation $\Delta k . \Delta x \cong 1$ as that which we have already obtained above for a wave group. Both the derivations follow from one another by an exchange of the quantities x and k. From the corpuscular point of view, this exchange corresponds to the exchange of the position and velocity as the quantity to be primarily determined with a given accuracy.

We thus see that the simultaneous assignment to a material particle of a *definite position* and of a *definite velocity* (or momentum)—which is the usual way of defining its *state* in classical mechanics—is impossible in wave mechanics. If the velocity of the particle is exactly specified, then the particle is represented by a system of plane harmonic waves with the same amplitude throughout all space so that its position is

entirely indeterminate. If, on the other hand, we locate the particle at a definite point and so reduce the extension of the associated wave packet to zero, then the velocity of the particle becomes entirely indeterminate, because all values and directions of this velocity, as determined by the amplitudes of the plane harmonic waves constituting the packet, are equally probable.

The *state* of a material particle in the sense of the classical theory (simultaneous assignment of position and velocity) can thus be determined according to wave mechanics only with a certain limited degree of accuracy with respect both to position and velocity. The amount of the uncertainty, as measured by the product of the actual error in the determination of one of the quantities x or v_x by the resulting 'effective' error in the other, is approximately equal to h/m (actually it is larger than this).

This result can be expressed in a somewhat different way in terms of the phase space of the classical statistical mechanics. If the motion of the particle is limited to the one dimension x, the phase space reduces to a plane with rectangular coordinates x, v_x specifying the various possible states of motion. Each point in the plane specifies a classical state of motion. The product $\Delta x . \Delta v_x$ represents a certain (rectangular) area in the phase plane. Thus the uncertainty relation $\Delta x . \Delta v_x \cong h/m$ means that a state of motion specifiable in the sense of wave mechanics corresponds approximately to an area h/m of the phase plane, and so includes as possible issues all of the exactly specifiable states of motion represented by the various points lying within this area.

Replacing the motion in one dimension by a three-dimensional motion we get instead of the phase plane a six-dimensional 'phase space' with the rectangular coordinates x, y, z, v_x, v_y, v_z and a phase volume $(h/m)^3$ for a specifiable state of motion. If, as is usually done, the velocity is replaced by the momentum $\mathfrak{g} = m\mathbf{v}$, then the *elementary volume* of the phase space corresponding to the effective uncertainty in the specification of the state of motion according to wave mechanics becomes equal to h^3. It should be remarked that the shape of this volume, so long at least as it remains rectangular, can be chosen at random.

In classical mechanics the state of motion not only can but *must* be specified by assigning simultaneously both the position and the velocity, that is, by assigning the exact values of the coordinates x, y, z and of the velocity components v_x, v_y, v_z. This follows from the character

of the classical equations of motion

$$m\frac{dv_x}{dt} = F_x, \qquad m\frac{dv_y}{dt} = F_y, \qquad m\frac{dv_z}{dt} = F_z,$$

where the components of the force $\mathbf{F}$ are defined as certain functions of the coordinates. Thus if the values of the coordinates *and* of the velocity components for some initial instant $t = 0$ are exactly known, then, and only then, can we determine them for any other instant $t \neq 0$. In other words, the 'state' of the particle is uniquely determined for all times by its state at any particular instant—and of course by the field of force $\mathbf{F}$.

Turning to wave mechanics, we meet an entirely different situation both with respect to the definition of the state and with respect to the equations of motion. The definition of a state, of course, must correspond to the equations of motion, and these, in wave mechanics, must not directly and exactly determine the motion of the particle itself, but rather the wave function $\psi(x, y, z, t)$ which gives the probability of finding the particle at any place at a given instant. If we commence by specifying the state of the particle at $t = 0$ by some initial form of the associated wave packet, we must then follow the course of the particle by determining the evolution of this packet.

It is natural to think *a priori* that the effective phase extension corresponding to this wave packet would remain constant—in agreement with the wave-mechanical uncertainty relations (26) on the one hand and with Liouville's well-known theorem of classical statistical mechanics on the other. (According to Liouville's theorem the volume of the phase space containing the state of a moving particle whose initial state was not exactly specified remains constant, i.e. independent of the time.)

In the case of an 'elementary' wave group defined by equation (22), § 7, with $a_\nu = \text{const.}$ inside and $a_\nu = 0$ outside $\Delta\nu$, both Δk and Δx are independent of the time—the former by definition and the latter according to (22a). If, however, Δk is not sharply defined, Δx in general varies with the time, so that the product of the effective values of Δk and Δx increases with $|t|$. This modification of the wave packet is described as *spreading* or *diffusion*.

For the sake of illustration we shall consider a non-elementary wave group defined by the equation

$$\psi(x, t) = \int_{-\infty}^{+\infty} a(k) e^{i2\pi(kx - \nu t)}\, dk, \tag{28}$$

with

$$a(k) = e^{-(k-k_0)^2/q^2}. \tag{28a}$$

The dependence of the amplitude of the constituent plane harmonic waves on the wave number is here represented by a Gaussian curve with its maximum at $k = k_0$ and an effective width $q = \Delta k$.

Let us first consider the shape of the packet, that is, the dependence of ψ upon x, at the initial instant $t = 0$. Putting $(k-k_0)/q = \xi$, we have

$$\psi(x, 0) = e^{i2\pi k_0 x} q \int_{-\infty}^{+\infty} e^{-\xi^2 + i2\pi q x \xi}\, d\xi,$$

or since

$$-\xi^2 + i2\pi q x \xi = -(\xi - i\pi q x)^2 - \pi^2 q^2 x^2 = -\xi'^2 - \pi^2 q^2 x^2,$$

and

$$d\xi' = d\xi,$$

we obtain

$$\psi(x, 0) = q e^{i2\pi k_0 x} e^{-\pi^2 q^2 x^2} \int_{-\infty}^{+\infty} e^{-\xi'^2}\, d\xi',$$

$$= \sqrt{\pi}\, q e^{-\pi^2 q^2 x^2} e^{i2\pi k_0 x}.$$

We thus see that at the initial instant the wave packet can be represented as a train of harmonic waves with the wave number k_0 and a variable amplitude

$$A(x) = \sqrt{\pi}\, q e^{-\pi^2 q^2 x^2}, \tag{28b}$$

whose dependence upon x is similar to that of $a(k)$ upon k. It will be remembered that A^2 is the measure of the probability of finding the associated particle at the corresponding point. The maximum of this probability is located at $x = 0$, and the effective width p of the probability packet, as defined by the formula $A(x) \sim e^{-x^2/p^2}$, is given by the relation $\Delta x = p = 1/\pi q$, which is practically equivalent to $\Delta k . \Delta x \cong 1$.

To determine the shape of the wave group, that is, the function $\psi(x, t)$, for $t \neq 0$ according to (28), we must define the frequency ν as a function of k (the approximate relation $\nu = \nu_0 + \left(\frac{d\nu}{dk}\right)_0 (k-k_0)$ being now insufficient). The exact relation between ν and k is [cf. eq. (18)]

$$\nu = c\sqrt{\left(\frac{m_0^2 c^2}{h^2} + k^2\right)}. \tag{29}$$

Assuming, for the sake of simplicity, that k is never very large, we get approximately

$$\nu = \frac{m_0 c^2}{h}\left(1 + \frac{h^2 k^2}{2m_0^2 c^2}\right) = \frac{m_0 c^2}{h} + \frac{hk^2}{2m_0}. \tag{29a}$$

This relation corresponds to the non-relativistic equation connecting the energy $\epsilon = h\nu$ (including the rest energy $m_0 c^2$) with the momentum $g = hk$, that is,

$$\epsilon = m_0 c^2 + \tfrac{1}{2} g^2/m_0.$$

Substituting (29 a) into (28) and putting $k = k_0+q\xi$ and

$$\nu_0 = \frac{m_0 c^2}{h}+\frac{hk_0^2}{2m_0},$$

we obtain

$$\psi(x,t) = qe^{i2\pi(k_0 x-\nu_0 t)}\int_{-\infty}^{+\infty} e^{-\xi^2+i2\pi\left(qx\xi-\frac{hk_0 q\xi t}{m_0}-\frac{hq^2\xi^2 t}{2m_0}\right)}\,d\xi.$$

The exponent in the integrand can be written in the form

$$-\alpha\xi^2+2i\beta\xi = -\alpha\left(\xi-i\frac{\beta}{\alpha}\right)^2-\frac{\beta^2}{\alpha},$$

with

$$\alpha = 1+i\frac{\pi h}{m_0}q^2t, \qquad \beta = \pi q\left(x-\frac{hk_0}{m_0}t\right) = \pi q(x-v_0 t).$$

Here $v_0 = hk_0/m_0$ is the average velocity of the particle associated with the wave group. Putting $\xi-i\beta/\alpha = \xi'$, we get, as before,

$$\psi(x,t) = \frac{\sqrt{\pi}}{\sqrt{\alpha}}q\,e^{-\beta^2/\alpha}\,e^{i2\pi(k_0 x-\nu_0 t)},$$

that is, a quasi-harmonic wave train k_0, ν_0 with a variable amplitude $A(x,t) = \frac{\sqrt{\pi}}{\sqrt{\alpha}}q\,e^{-\beta^2/\alpha}$. Since α is a complex quantity this expression is strictly speaking not the real amplitude which we have considered hitherto, but is its product by a factor of the form $e^{i\phi}$, the phase ϕ being also a certain function of x and t.

The real amplitude is equal to the *modulus* of the 'complex amplitude' $A(x,t)$ and its square is equal to the product of A and the conjugate complex quantity A^*. Since the modulus of $e^{i2\pi(k_0 x-\nu_0 t)}$ is equal to 1, the square of the real amplitude is equal to the product of ψ and the conjugate complex function ψ^*. We thus obtain

$$|A|^2 = \psi\psi^* = \frac{\pi q^2}{\sqrt{(\alpha\alpha^*)}}\,e^{-\beta^2\left(\frac{1}{\alpha}+\frac{1}{\alpha^*}\right)},$$

that is

$$|A|^2 = \frac{\pi q^2}{\sqrt{\left\{1+\left(\frac{\pi\hbar q^2 t}{m_0}\right)^2\right\}}}\,e^{-\frac{2\pi^2 q^2(x-v_0 t)^2}{1+(\pi\hbar q^2 t/m_0)^2}}. \tag{30}$$

This formula shows that the distribution of probability in the wave packet is again represented by a Gaussian curve with its maximum at the point $x = v_0 t$ moving with the average velocity of the particle and having an effective width p_t determined by the relation

$$p_t = \frac{1}{q\pi}\sqrt{\left\{1+\left(\frac{\pi hq^2 t}{m_0}\right)^2\right\}}, \tag{30a}$$

or

$$p_t = \sqrt{\left\{p^2+\left(\frac{hqt}{m_0}\right)^2\right\}}, \tag{30b}$$

$p = 1/(\pi q)$ being the initial width of the wave packet. p_t is defined as before by putting $|A| \sim e^{-(x-v_0 t)^2/p_t^2}$.

The preceding formula, obtained by C. G. Darwin, shows that the width of the (not sharply defined) wave packet under consideration varies like the hypotenuse of a right-angled triangle with one side constant and equal to p, and the other side equal to $hq|t|/m_0$.

The spreading of the packet proceeds therefore—unlike the diffusion of a swarm of particles—symmetrically with regard to the time (or more exactly with regard to the initial instant $t = 0$), which means that it can be considered as a *reversible* process.

Another remarkable feature of the relation (30 b) is that not only the initial breadth of the packet but also its *rate of spreading* is determined by (30 b) in conformity with the uncertainty relation $\Delta x . \Delta v \cong h/m_0$, and, indeed, this can be considered as a new derivation of the relation.

The cause of spreading lies, from the corpuscular point of view, in the inaccurate determination of the velocity of the particle, which increases the uncertainty of its position as we proceed from the initial moment forward into the future—or backward into the past. If we denote the inaccuracy in the (original) velocity determination by $\pm\Delta v$ we can represent the increase in the breadth of the packet in time t by the product $\Delta v . |t|$. Equating this to the expression $\pm hqt/m_0$ in (30 b), we get the relation $\Delta v = \pm kq/m_0$. This agrees with the relation $v = hk/m_0$ or $\Delta v = h . \Delta k/m_0$, since $\Delta k = \pm q$ is merely the original effective inaccuracy in the determination of the wave number k.

For large values of $|t|$ formula (30 b) reduces to

$$p = \pm \frac{hq}{m_0} t = \Delta v . t.$$

The fact that for smaller values of t the rate of change of p is not constant can be interpreted as due to ambiguity in the sign of Δv.

The preceding analysis therefore shows that the essence of the difference between classical mechanics and the new wave or 'probability' mechanics may be traced to the fact that the latter recognizes the impossibility of an exact determination of the state of motion of a particle, that is, of the simultaneous specification of its coordinates and their *rates of change*, the corresponding velocity components. This point of view was developed by Heisenberg in 1927 as a key to the understanding of the nature of wave mechanics. It seems to me, however, that this key is provided not by Heisenberg's principle 'of the fundamental unobservability of conjugate quantities' such as a coordinate and its

time derivative—that is, of the impossibility of their simultaneous exact determination—but rather by Born's probability concept, which forms the basis of the wave-corpuscular parallelism.

Heisenberg's 'uncertainty relations' $\Delta x.\Delta g_x \cong h$ or $\Delta x.\Delta v_x = h/m$, etc., however, show very clearly why the old deterministic mechanics is so successful in the realm of ordinary macroscopic particles and fails only when applied to the elementary particles of matter, such as electrons, having a *very small mass*. Thus for a particle with a mass equal to one gramme the relation $\Delta x.\Delta v_x = \hbar/m$ gives $\Delta x.\Delta v_x \cong 10^{-27}$. Assuming the minimum inaccuracy in the experimental location of the particle to be $\Delta x = 10^{-5}$ cm. we get for the corresponding limiting inaccuracy in the determination of its velocity a value $\Delta v = 10^{-22}$ cm./sec. The 'kinematic inaccuracy' connected with the motion of macroscopic particles is thus quite negligible. The situation is entirely different in the case of electrons. For the latter $m = 9 \times 10^{-28}$, and consequently $\Delta x.\Delta v_x \cong 1$. With $\Delta x \cong 10^{-13}$ (the classical 'radius' of the electron, given by the electromagnetic theory of mass) this gives $\Delta v_x \cong 10^{13}$ cm./sec. Therefore, in attempting to follow experimentally the motion of an elementary particle, we come upon a difficulty, summed up in the uncertainty relations, which does not manifest itself practically in the case of ordinary macroscopic bodies (considered as point-like material particles).

10. Heisenberg's Theory of the Uncertainty Relation and the Role of the Observer

The derivation of the uncertainty relations from the theory of wave packets, given above, was indicated by Bohr. Heisenberg derived these relations, as well as the fundamental 'principle of unobservability' underlying them, from entirely different considerations, based not on the analysis of the probability connexion between waves and particles, but on a study of the experimental conditions under which these particles are—or may be—observed, with the incidental use of the wave-corpuscular parallelism, but—as it seems to me—without sufficient appreciation of its importance.

Heisenberg's argument is illustrated by the following analysis of the conditions under which the position and the motion of an elementary particle—an electron, say—can be traced experimentally.

The position of an electron (or any other very small particle of matter) can be determined—at least in principle—by observation of the light rays reflected, scattered or emitted by it. The particle will

behave in this case like a point-source of light; and such a point-source observed with the eye or with a microscope appears as a small *diffraction circle* whose radius is approximately equal to the wave-length λ of the light. Therefore the position of the particle can be determined only with a limited accuracy—of the order of magnitude of λ.

Now *in principle* we can choose λ arbitrarily small and can observe the particle, not with ordinary light, but with gamma-rays (the practical impossibility of constructing a γ-ray microscope does not concern us). If the particle is an electron, then in attempting to determine its position by observation of short wave-length light rays scattered by it, we obtain an elementary *Compton effect* by which the momentum of the electron *during*, or better still, *in consequence* of the observation, alters by an amount of the order of magnitude of h/λ. We obtain the same result if we try to locate an atom, not by scattered light, but by the light that it emits itself. If the light is emitted in the form of a quantum $h\nu$, the atom must experience a recoil of the amount h/λ, so that again the observation of the atom's position is connected with an alteration of its momentum (velocity). In both cases the accuracy of location of the particle by means of the scattered or emitted light is equal to the wave-length of the light, and the associated change of momentum is inversely proportional to this wave-length, so that by increasing the accuracy in the determination of the particle's position we increase the accompanying change of its momentum.

Now the important point about this change is that it cannot be followed in detail and described as a continuous variation of the momentum with time. The change of momentum associated with the scattering or emission of a light quantum has to be visualized as an *instantaneous transition* from the initial to the final value. It thus becomes impossible to ascertain the value of the particle's momentum at *the instant* of time for which its position is being determined, for just at this instant the value is making an abrupt jump. The magnitude of this jump can obviously be considered as a measure of the uncertainty in the value of the momentum corresponding to the observed position and can be regarded as *produced by the* observation. Since the position itself can be observed with only a limited accuracy of the order of magnitude λ and the corresponding uncertainty in the value of the momentum is of the order of magnitude of h/λ, the product of the two inaccuracies—or rather of the *inaccuracy* of the observation of the coordinates of the particle and the resulting *uncertainty* in the evaluation of its momentum *for the instant of observation*—is approximately

equal to h. This agrees with the uncertainty relation $\Delta x . \Delta g_x \cong h$ obtained above from the analysis of the structure of a wave packet associated with the motion of a particle.

A similar argument applies to an observation or experiment tending to ascertain the velocity or the momentum of an elementary particle. It can be shown, with the help of de Broglie's relation between velocity and wave-length, that the use of any device for the determination of this momentum introduces an uncertainty in the *simultaneous* position of the particle, this uncertainty being inversely proportional to the accuracy with which the measurement of momentum is carried out.

Heisenberg's derivation of the uncertainty relations thus brings forth a new principle, entirely alien to classical mechanics—namely, that it is impossible to *measure simultaneously* with any degree of precision the position and the momentum of a material particle, because the measurement of the one implies, or produces, an uncertainty in the estimation of the other. Heisenberg himself and many other writers and commentators expressed the belief that this principle—often referred to as the principle of fundamental unobservability (*prinzipielle Unbeobachtbarkeit*) is to be considered as the basic principle of the new 'quantum' mechanics—similar in character and scope to the principle of relativity underlying Einstein's relativity theory. In both cases the observer becomes intimately connected with the things or phenomena observed, so that they cannot be described without any reference to the observer. In the relativity theory events must be referred to a system of co-ordinates connected with the observer. In the quantum theory one must take into account the unavoidable alterations of the phenomena produced by the observer's interference. This excludes the possibility of describing these phenomena as they would have been without this interference.

The alterations referred to are very small and unimportant in the case of big things (ordinary bodies) and macroscopic phenomena, and so these, and in particular the phenomena of the motion of ordinary bodies, can be described in a perfectly deterministic way, excluding any uncertainty whatsoever and ignoring the active role of the observer.

The situation is entirely different in the case of elementary particles and microscopic phenomena. Here the alterations connected with their observation (i.e. produced by the measurement of some of the quantities characterizing them) become relatively large and important. Therefore the description of the phenomena *'an sich'—such as they would appear*

if they had not been observed and modified in an unascertainable way by the observation—becomes impossible.

If, then, we wish to build up a theory which will *adequately* describe and predict microscopic phenomena, that is, in conformity with the results that can be obtained from and verified by experiments or observations, it must incorporate the uncertainties which are inherent in such experiments or observations.

The experimenter or observer must be dealt with, not as the passive contemplative element of the classical theory, but as an active disturbing factor. In eliminating this factor from our description we must take into account the uncertainties produced by it. This means that a deterministic theory of microscopic phenomena—in particular a deterministic *mechanics* dealing with the motion of elementary particles—is neither necessary nor possible.

I have presented Heisenberg's theory of the uncertainty relations and his principle of the observer's interference at such length, not because I consider them to be so fundamental, but on the contrary because I think that their importance has been greatly exaggerated (especially in popular and philosophical discussions of the new mechanics) and their meaning misunderstood.

First of all let us remark that the principle of the observer's interference in its *general* form—'the impossibility of observing a phenomenon without altering it'—is very old and in fact quite trivial. Every physicist will admit that the measurement of the temperature or electric potential of a body or of any other physical quantity must alter it to some extent. Usually the measuring device is made so small compared with the body investigated that this alteration is negligible. But in principle it is always there.

Now until recently this circumstance did not in the least undermine our belief in the deterministic character of the phenomena studied. The reason for this clearly lies in the fact that the changes produced by the observation (measurement) in the macroscopical phenomena under investigation can always be exactly accounted for and theoretically eliminated along with the observer responsible for them. For instance, the change of the temperature or the electric potential of a body produced by connecting it with a thermometer or an electrometer can be easily calculated from the knowledge of the heat capacity of the former or of the electric capacity of the latter. By introducing the corresponding correction for the temperature or the potential, we can determine what they would have been without our interference. We thus

see that Heisenberg's principle—the impossibility of observing a phenomenon without altering it—does not challenge determinism so long as the alterations implied are themselves controllable, that is, deterministic.

The application of this principle to the phenomena of the motion of elementary material particles does not lead to indeterminism—as is often believed—but *simply reveals the indeterminism which is pre-existing in these phenomena* and which consists in the impossibility of ascertaining the exact position and the exact velocity of the particle at the same moment. Heisenberg's argument concerning the investigation of the particle's motion by means of scattered or emitted light describes the situation from the point of view of the experimental physicist; the argument given before from the wave packet theory presents the matter from a purely theoretical standpoint, ignoring the experimenter. The prominence of the latter in Heisenberg's discussion does not seem to be an essential feature of the situation and does not at all correspond to the difference between the old deterministic and the new 'indeterministic' mechanics. As pointed out above, this difference is due to the different type of variation in time of those alterations which are connected with observation and measurement according to the classical and the new views, and not to the fact that these alterations were neglected in the one case and taken into account in the other.

The situation is not substantially changed by the fact that, in Heisenberg's application of the 'observer's interference' principle, the observation does not alter the quantity which is directly observed (e.g. the x-coordinate) but another quantity called its conjugate (the x-component of the momentum). As a matter of fact we meet with an entirely similar situation in classical theory. In the measurement of temperature or electric potential, for example, the quantity directly affected is the thermal energy or the electrical charge, i.e. the quantity dynamically conjugate to that which is being measured. The uncertainty in the conjugate quantity in Heisenberg's treatment must be ascribed to the indeterminate character of its variation with the time (indicated by the word 'transition' or 'jump'), which makes it impossible to ascertain its value for the instant of time corresponding to a given value of the quantity directly observed.—In the treatment by wave packets the uncertainties of the two conjugate quantities are both dealt with on the same basis, both being due to the impossibility of a sharp definition of a '*state*' in the sense of classical mechanics.

The exaggeration of the role of the observer in the study of elementary

phenomena has led to a great deal of confusion and misunderstanding. It has given rise, for instance, to the common popular opinion that the size of the probability packet representing the motion of a particle suddenly shrinks when this particle, after moving for a while unperturbed, is located by means of the experiment with scattered light described above, the shrinking being due to the action of the light. This conception is completely wrong. The shrinking of the probability packet connected with the more exact location of the particle is not a *physical* process which takes place because of the observation, but a *logical* process, undertaken deliberately by the observer, and consisting in a new estimation of the probability of some intermediary state of the process studied. Such arbitrary re-estimations are always made, even in the classical theory, when an intermediate state is taken as the starting-point of a new development. Equally misleading is the expression, which is becoming current, that the observer 'prepares' the system which is to be considered. This 'preparation' is again a logical process and not a physical process produced by the observation.

III

WAVE MECHANICS OF THE MOTION OF A PARTICLE IN A FIELD OF FORCE

11. Bohr's Quantum Theory of the Atom and Standing Cathode Waves

So far we have entirely neglected the quantum theory of the atom formulated by Bohr in 1913. Yet it was in this field that the wave theory of matter achieved its first victory.

It should be remembered that Bohr's theory, with its basic idea of *stationary states* of the atom, as well as of other mechanical systems, arose historically in conjunction with the quantum theory of light. From the fact that atoms lose or absorb energy by radiation only in definite quanta of finite amounts, Bohr concluded that the processes of emission and absorption are discrete elementary acts which carry the atom over from one radiationless state into another state which is likewise radiationless. To these radiationless, steady, or 'quantized' states of motion belong definite discrete values of the energy $W_1, W_2, W_3, \ldots$. By a transition from the nth to the mth state the energy difference $W_n - W_m$ is emitted as monochromatic light of frequency

$$\nu = \frac{W_n - W_m}{h}. \tag{31}$$

Bohr left open the question of the structure of light. Einstein showed later (1917) that Bohr's mechanism of light emission and absorption contradicts the usual wave theory of light and is in complete agreement with the corpuscular theory. By considering the heat equilibrium between atoms and radiation, Einstein in fact was able to show that on emission or absorption the atom, in addition to an alteration of energy $W_n - W_m = h\nu$, must also experience an 'impact', that is, an alteration of its momentum by the amount $h\nu/c = h/\lambda$.

Further, Bohr set up the simplest 'quantum conditions', those which determine the 'quantized' states of the hydrogen atom. If we consider only circular orbits of the electron about the nucleus (supposed to be at rest), then Bohr's original quantum condition runs

$$rmv = n\frac{h}{2\pi}, \tag{32}$$

where r is the radius of the orbit, mv the momentum of the electron, and n is a positive integer.

This quantum condition was later generalized by Sommerfeld and written in the form

$$\oint p_1\,dq_1 = n_1 h, \qquad \oint p_2\,dq_2 = n_2 h,\ldots, \qquad \oint p_f\,dq_f = n_f h. \tag{32a}$$

q_1, q_2,... are the so-called 'separable' coordinates of the given mechanical system, i.e. generalized coordinates such that the 'momenta' p_1, p_2,... can be expressed as functions of the corresponding coordinates alone. The momenta are defined in general by the formulae $p_k = \partial T/\partial \dot{q}_k$, where T is the kinetic energy of the system expressed as a function of the velocities $\dot{q}_k = dq_k/dt$ and the coordinates q_k. Even when 'separable coordinates' exist the quantum conditions (32a) can only be formulated if the functions $p_n(q_n)$ are periodic or if they only permit values of the argument lying within certain limits. The integrations in (32a) are to be carried out for these periods or for the to-and-fro oscillation of the argument between the limits concerned.

No rational interpretation of the Bohr-Sommerfeld quantum conditions could be given in terms of the ordinary corpuscular conceptions of the nature of matter. Before the appearance of de Broglie's theory they were only rules found in a semi-empirical way and supported by their agreement with the facts of experiment. Ehrenfest was able later in a certain measure to justify them by means of 'the principle of adiabatic invariance' and Bohr by his 'correspondence principle'. By refining the correspondence principle, Heisenberg, in 1925, was able to formulate the exact mathematical theory of micromechanical processes —a formulation which was soon shown to be formally equivalent to the de Broglie-Schrödinger wave mechanics which arose almost simultaneously with it, but nevertheless, independently of it (see part II). However, the way taken by Heisenberg was a *detour*. The direct way to an understanding of the quantum laws of atomic mechanics was pointed out by de Broglie with his idea of phase waves—an idea that was at first very obscure.

If we introduce the de Broglie wave-length $\lambda = h/(mv)$ corresponding to the momentum $g = mv$ into the Bohr quantum condition (32), then we see that $2\pi r = n\lambda$ and the quantum condition assumes the following very simple and vivid form: *in a quantized state of motion of the hydrogen atom the length of the circular electron orbit* $(2\pi r)$ *is equal to an integral multiple of the corresponding wave-length.*

This interpretation can easily be extended to include closed orbits of a more general form, e.g. elliptical orbits. If we add the equations $\oint p_1\,dq_1 = n_1 h$ and $\oint p_2\,dq_2 = n_2 h$ which correspond to a periodic Kepler

motion of the electron about a positive nucleus, we get, on taking into account the relation $p_1\dot{q}_1+p_2\dot{q}_2 = 2T$ (T = kinetic energy):

$$\oint (p_1\,dq_1 + p_2\,dq_2) = \oint 2T\,dt = \oint mv^2\,dt = \oint mv\,ds,$$

where $ds = vdt$ is the element of the arc of the orbit. If we replace mv by h/λ, we get:

$$\oint \frac{ds}{\lambda} = n \qquad (= n_1+n_2). \tag{32b}$$

In this integral the wave-length is to be regarded as a variable quantity which depends on the position of the electron in its orbit.

The interpretation of the quantum conditions given by de Broglie was the first, as it were, 'empirical' argument in favour of his idea of phase waves. The possibility of carrying out this idea by transferring Einstein's light quanta relations $\epsilon = h\nu$ and $g = h/\lambda$ to corpuscles with non-vanishing mass appeared at the time to be very questionable, especially with the 'naïve' assumption of the relation between waves and corpuscles which was suggested by de Broglie. This idea gave rise to considerable confusion with regard to the interpretation just mentioned. De Broglie defined his wave-lengths for circular electron orbits by the formula $2\pi r = n\lambda$ only for points which lie on the orbit. He left unanswered the question of the length and general character of the wave processes *outside* the quantized orbit described by the electron. The existence of such wave processes in the surrounding space was scarcely taken into consideration.

Two years later the fundamental ideas of de Broglie's theory were developed and extended by the work of Schrödinger. Schrödinger was at first influenced by the false idea of the wave packet as an actual substitute for the particle of the classical theory. This idea, however, proved to be very fruitful, for it led him to investigate the laws of the propagation of the de Broglie waves *as such*, independently of the particles connected with them. The result of this investigation was the discovery of the 'wave equation' which determines the various states of vibration which are possible in a given external field of force.

We can understand Bohr's quantization principle, however, without the use of this equation. For in the simplest case of the free motion of a particle the associated wave system can be considered as directly known and indeed can be represented by the superposition of a number of 'elementary' wave systems, that is, plane harmonic wave systems with arbitrary amplitudes, frequencies, and directions of propagation. So far we have limited ourselves to such elementary wave systems

and to those combined systems which form wave groups or wave packets. These wave systems could be associated with the motion of a *completely* free particle through *all space*. If, however, we wish to interpret Bohr's quantized states, we must consider electrons which are bound by some field of force to stay within a limited region of space.

The simplest example of such a field of force is that provided by a force-free region of space V, surrounded by a closed surface S consisting of a concentrated surface field of force of infinite intensity, directed normally to the surface. This acts like an impenetrable barrier and so prevents the electron in V from escaping to the outside. Thus such a 'bound' electron can be pictured as enclosed in a 'box', within which its potential energy is equal to zero and outside is equal to infinity, the walls of the box being infinitely thin.

The behaviour of such an electron which is free to move within a limited region of space V can be described wave-mechanically by means of a wave system of the same type as that associated with a completely free electron, with the restriction that the *amplitude of the vibrations must vanish at the boundary of the region considered*. This means, from the point of view of the probability interpretation of the wave-corpuscle parallelism, that the electron is unable to cross the limiting surface S (this would also remain true if the stopping field of force at S were actually absent). Of course we are concerned only with that portion of the resulting wave system which is contained within S, the rest having no physical meaning. Such a limited wave system, having a *nodal surface* at the boundary, is called a *standing* wave system.

The simplest example of a standing wave system is obtained by superposing two elementary plane harmonic waves with the same amplitude and frequency but travelling in opposite directions. Such a wave system can be associated with an electron (or any other particle) moving parallel to the x-axis either in the positive or in the negative direction, between two infinite plane walls perpendicular to x at a certain distance a from each other.

The theory of standing waves of this type is exactly similar to the theory of the standing waves—or in other words the free vibrations—of a uniform string of length a. In the case of a string we distinguish the fundamental vibration, which has no nodes except those at the (fixed) ends of the string, and the overtones, which have a certain number of equidistant nodes, that is, non-vibrating points, along the string. These so-called 'normal' or 'simple' vibrations of the string can

be superimposed with arbitrary amplitudes, giving rise to a 'compounded' vibration of a more complicated type.

The distance between the successive nodes in a simple standing wave system is equal, as is well known, to half the wave-length of the two corresponding elementary travelling waves. The wave-length corresponding to the nth normal mode of vibration of a string, for which the nodal distance is equal to a/n, is thus given by

$$\lambda_n = 2a/n. \tag{33}$$

This formula applies equally well to the standing cathode waves or normal 'cathode' vibrations associated with the to-and-fro motion of an electron between the two walls separated by a distance a. Introducing the momentum of the electron $g = h/\lambda$, we see that to the nth normal mode of vibration there corresponds a state of motion with momentum

$$g_n = \frac{hn}{2a}, \tag{33 a}$$

and (kinetic) energy

$$W_n = \frac{1}{2m} g_n^2 = \frac{h^2 n^2}{8ma^2}. \tag{33 b}$$

Since there are no other possible (simple) modes of vibrations, the above-defined states of motion must be the only ones that can occur in nature. They are obviously the 'steady' or 'quantized' states of Bohr's theory. In fact the relations (33 a) are easily seen to be a special case of the Bohr-Sommerfeld quantum conditions (32 a) applied to the problem considered. Thus, if we put $q = x$ and $W = \frac{1}{2}m\dot{q}^2$, we get $p = m\dot{q} = g$. The integration indicated by $\oint$ must be extended from $x = 0$ to $x = a$ and back to $x = 0$. Thus

$$\oint p\, dq = \int_0^a g\, dx + \int_a^0 g\, dx = |g| \int_0^a dx - |g| \int_a^0 dx.$$

This gives

$$2a|g| = nh,$$

which is identical with (33 a), g_n being the absolute value of g in the nth steady state. We see further that the quantum number n of the Bohr theory can be interpreted wave-mechanically as the *number of nodes* in the normal vibration associated with the corresponding state. It may be mentioned that the fundamental vibration with $n = 1$ corresponds to the steady state of the smallest energy, usually referred to as the 'normal state', whereas the 'overtones' ($n > 1$) correspond to the so-called 'excited states'.

The above example serves to illustrate the difference between the

correct mechanical theory of the motion of a bound electron, based upon the probability interpretation of the wave-corpuscle parallelism, and the original, incorrect interpretation of this parallelism by de Broglie. In the case of an electron moving to and fro between two perfectly reflecting walls, the 'orbit' of the electron, according to the classical theory, is a double rectilinear segment of total length $2a$— corresponding to the circular orbit of length $2\pi r$ in the case of a hydrogen atom. Applying de Broglie's principle that the length of the orbit is equal to an integral multiple of the wave-length $\lambda = h/|g|$ to the case under consideration, we get $2a = nh/|g|$ or $|g| = nh/2a$ in accordance with (33 a).

Although this result is correct, yet the implication that the motion of the electron takes place and can be followed in a classical way, the electron having a definite position and momentum ($+|g|$ or $-|g|$) at any given instant, is erroneous. As a matter of fact there is no such thing as the 'orbit' of the electron. Both the position and the momentum or velocity (including direction) of the electron at any instant t remain indeterminate. All that can be determined is the *probability* of finding the electron within a certain range of its region of motion, between x and $x+dx$ say, and the probability of its moving in the positive or the negative direction, the magnitude of the velocity being fixed, that is, exactly known, for each steady state.

It may seem at first sight that the exact knowledge of the magnitude of the electron's velocity is inconsistent, from the point of view of the uncertainty principle discussed above, with the fact that its position is fixed with a finite degree of accuracy within a given segment a of the x-axis. One might think, for instance, that a decrease of a should involve an increasing uncertainty in the velocity of the electron, not only with regard to its direction but also with regard to its amplitude. On the other hand, the preceding theory shows that this magnitude remains exactly fixed (for each steady state), however small a may be.

It can easily be seen that the whole argument upon which the derivation of the uncertainty relations in § 9 was based, is not applicable to the present case, i.e. to a *bound* electron (or any other particle), and that these relations must be modified in such a way as to allow for the *discontinuous* character of the series, formed by the values of the parameter k, defining the velocity or the energy of the particle in various steady states. For a *free* particle these values form a continuous sequence.

In fact a wave group of the form $\int a_k \psi_k \, dk$ where the integration is

extended over a small range of the variable k, cannot be used to describe the motion of a particle enclosed within a finite range a of the x-axis. We can replace it, however, by a wave group of a somewhat different type, formed by the superposition of a number of simple standing waves

$$\psi_n = A_n \sin\frac{\pi n}{a} x \, e^{-i2\pi\nu_n t} \qquad (n = 1, 2, \ldots) \tag{34}$$

with arbitrary amplitudes A_n. The latter can be chosen in such a way that at a particular instant, $t = 0$ say, the function $\psi = \sum \psi_n$ reduces to any given function of x, vanishing at the ends of the interval a. In particular $\psi_{t=0}$ can reduce to a function vanishing outside a small part Δa of a, and thus representing a particle which *at the initial instant* $t = 0$ is supposed to be located within Δa. For $t \neq 0$ the function ψ will, however, have an entirely different form, corresponding to a finite probability of finding the particle at any point of a outside Δa.

We shall not consider this question (which amounts to the expansion of the function $\psi_{t=0}$ in a Fourier series) at any greater length. All we are concerned with here is this: *firstly*, any localization of the particle within the allowed interval a involves a definite instant of time, for which this localization is introduced, and *secondly*, it requires the introduction of a *number* of steady states *with different energies* W_n, so that the value of the energy and consequently the magnitude of the velocity, associated with the particle, remains uncertain or indeterminate.

The amount of this uncertainty can be estimated from the difference of the quantized values of W_n or ν_n, etc., for two consecutive steady states, this difference representing in fact the *smallest* amount of uncertainty involved. It is more convenient, however, to use instead of $\Delta W_n = W_{n+1} - W_n$ or $\Delta\nu_n = \nu_{n+1} - \nu_n$ the area of the phase-plane (x, g), containing all the classical states which are enclosed between two consecutive quantized states. On the (x, g) diagram the classical states, associated with a given (nth) quantized state, are represented by the points of two segments $A'_nB'_n$ and $A''_nB''_n$ defined by the conditions $0 \leqslant x \leqslant a$ and $g = \pm g_n$ (Fig. 3).

The area between the two lines $A''_nB''_n$ and $A''_{n+1}B''_{n+1}$ is equal to $(g_{n+1} - g_n)a = \frac{1}{2}h$ according to (33 a), so that the total area between the nth and $(n+1)$th state is exactly equal to h. This result can be derived directly from the Bohr-Sommerfeld form of the quantum conditions, $\oint g\,dx = nh$, if one remembers that this integral gives the area of the rectangle $A'_nA''_nB''_nB'_n$. It can be interpreted in a slightly different manner—namely, that one quantized state corresponds to a number of

classical states filling a phase extension h. On Fig. 3 the classical states, corresponding to the nth quantum state, are represented by two strips limited by the dotted lines running midway between the full ones.

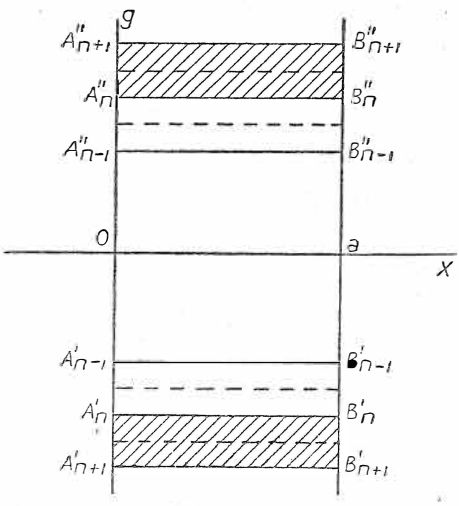

FIG. 3.

It follows from the preceding considerations that for a bound particle the uncertainty as to the position and momentum, apart from the steady states of motion, must be equal to or in general *larger than* h, if measured by the phase extension (x, g) of the classical theory.

Returning to the question of the probability of finding the electron in a given position for a steady state, we meet the following alternatives. The electron can on the one hand be associated, at any time t, with either *one* of the two elementary waves

$$\psi_+ = \frac{1}{2i} A e^{i2\pi(kx-\nu t)}, \qquad \psi_- = -\frac{1}{2i} A e^{i2\pi(-kx-\nu t)}, \tag{34 a}$$

travelling respectively in the positive and the negative direction and forming, by their superposition, the standing wave system which represents the given steady state†

$$\psi = \psi_+ + \psi_- = A \sin 2\pi kx \, e^{-i2\pi\nu t}.$$

Alternatively it may be associated with this resulting wave system.

In the first case (which corresponds to the treatment of reflection and transmission outlined in § 12) the probability of finding the electron between x and $x+dx$ is equal to $|\psi_+|^2\,dx$ for the positive and $|\psi_-|^2\,dx$ for the negative direction of motion. The two probabilities are, of course, equal to one another ($|\psi_+|^2 = |\psi_-|^2 = \frac{1}{4}A^2$) and are independent of x.

In the second case the probability of finding the electron between x and $x+dx$ is equal to

$$|\psi|^2 dx = A^2 \sin^2 2\pi kx \, dx,$$

and is therefore a function of x which vanishes at the nodal points. The *average* value of the resulting probability density for different values of x, $\overline{|\psi|^2} = \frac{1}{2}A^2$, is equal to the sum of the probability densities

† The (complex) amplitudes of ψ_+ and ψ_- are chosen here in such a way that $\psi = 0$ for $x = 0$. The condition $\psi = 0$ for $x = a$ is satisfied if $k = 1/\lambda$ is determined in accordance with (33).

$|\psi_+|^2$ and $|\psi_-|^2$. This additive relation for the probabilities would be expected according to the classical probability theory, since the motion of the electron in the positive or in the negative direction are two alternative events, and the probability of any one of a number of such alternative events taking place is usually defined as the sum of the probabilities of each of these events taken separately. We thus see that the addition theorem of classical probability theory must be rejected and *replaced* by an addition theorem for the wave functions. In this form the addition theorem corresponds to the *superposition principle* which we have used until now as a matter of course, and of which equation (34 a) is a particular case.

Wave mechanics has been defined above, from the corpuscular point of view, as the probability theory of the motion of material particles. It must, however, be remembered that this probability theory, in so far as the addition theorem is concerned, is essentially different from classical probability theory. Another fundamental theorem of the classical probability theory—the multiplication theorem for the probabilities of independent events—can be shown to remain valid in the new probability theory with certain practically unimportant restrictions. This principle can be demonstrated—or at least illustrated—by generalizing the preceding results to apply to an electron (or any other particle) moving *freely* in three dimensions in an unlimited space or in a rectangular box.

In the following we shall drop the time factor $e^{-i2\pi\nu t}$ in the expression of the wave functions, since it plays no role in evaluating the probability so long as we are concerned with motions having *the same* resulting (total) *energy* $\epsilon = h\nu$. With this convention the free motion of an electron parallel to the axes x, y, z can be represented by the functions

$$\psi_x = A_x e^{i2\pi k_x x}, \qquad \psi_y = A_y e^{i2\pi k_y y}, \qquad \psi_z = A_z e^{i2\pi k_z z}, \tag{35}$$

the components of the momentum of the electron being equal respectively to

$$g_x = hk_x, \qquad g_y = hk_y, \qquad g_z = hk_z.$$

Let us now assume that these three independent motions take place simultaneously, so that the electron is actually moving in the direction of the resultant of the wave vectors k_x, k_y, k_z. Such a motion will be represented by a wave function

$$\psi = A e^{i2\pi(k_x x + k_y y + k_z z)}, \tag{35 a}$$

which can be defined as the *product* of the functions (35). The proba-

bility of finding the electron in the volume-element $dxdydz$ is proportional to

$$|\psi|^2\,dxdydz = A^2\,dxdydz.$$

From the point of view of the classical probability theory it must be equal to the product of the three factors

$$|\psi_x|^2\,dx = A_x^2\,dx, \qquad |\psi_y|^2\,dy = A_y^2\,dy, \qquad |\psi_z|^2\,dz = A_z^2\,dz$$

measuring the probabilities of finding the electron between x and $x+dx$ irrespective of y and z, between y and $y+dy$ irrespective of x and z, and between z and $z+dz$ irrespective of x and y. This result can be taken over into the new probability theory if we put $A = A_x.A_y.A_z$. In this case (35 a) reduces to

$$\psi = \psi_x.\psi_y.\psi_z, \tag{35 b}$$

which means that the multiplication theorem of the classical probability theory can be considered as a consequence of the corresponding theorem for the wave functions ('probability amplitudes') and can be incorporated without any change into the new probability theory.

The limitations of this result due to the omission of the time factor $e^{-i2\pi\nu t}$ will be considered later. For the present we wish to emphasize the fact that the validity of the multiplication theorem of the 'probability amplitudes' for the component motions of a particle depends *essentially* upon the circumstance that these probability amplitudes, that is, the wave functions ψ_x, ψ_y, ψ_z, are represented by the *complex* expressions (35) with imaginary exponentials. If we represented them by the *real* parts of these complex expressions—as is usually done in optics and acoustics—the multiplication theorem could not be maintained. This circumstance may serve as a first indication of the fact that the wave functions *must* be represented by complex expressions if they are to have the meaning of 'probability amplitudes'.

Let us now turn from the travelling waves (35) associated with a free electron to standing waves of the type

$$\psi_x = A_x \sin 2\pi k_x x, \qquad \psi_y = A_y \sin 2\pi k_y y, \qquad \psi_z = A_z \sin 2\pi k_z z, \tag{36}$$

obtained from (34 a) by omitting the time factor. These represent an electron moving respectively parallel to the x-axis between two planes $x = 0$ and $x = a$, parallel to the y-axis between the planes $y = 0$ and $y = b$ or parallel to the z-axis between the planes $z = 0$ and $z = c$. In order that the planes $x = a$, $y = b$, $z = c$ may be nodal planes—like the planes $x = 0$, $y = 0$, $z = 0$—we must put

$$k_x.a = \tfrac{1}{2}n_x, \qquad k_y.b = \tfrac{1}{2}n_y, \qquad k_z.c = \tfrac{1}{2}n_z, \tag{36 a}$$

n_x, n_y, n_z being arbitrary positive integers.

These three states of motion can obviously be combined to give the motion of an electron in a rectangular box with the edges a, b, c. If the multiplication theorem is actually valid, the wave function or probability amplitude ψ describing such a motion must be equal to the product of the wave functions (36).

Now replacing the sines by the sums of exponentials, that is, representing ψ_x as the superposition of two travelling waves $\psi_{x+} = \frac{1}{2i} A_x e^{i2\pi k_x x}$ and $\psi_{x-} = -\frac{1}{2i} A_x e^{-i2\pi k_x x}$, ψ_y as the sum of ψ_{y+} and ψ_{y-}, and ψ_z as the sum of ψ_{z+} and ψ_{z-}, we can represent the product

$$\psi = \psi_x . \psi_y . \psi_z = A \sin(2\pi k_x x) \sin(2\pi k_y y) \sin(2\pi k_z z) \qquad (37)$$

($A = A_x A_y A_z$) as a superposition (sum) of eight travelling waves of the type

$$\psi_{x\pm} \psi_{y\pm} \psi_{z\pm} = \pm \frac{A_x A_y A_z}{8i} e^{i2\pi(\pm k_x x \pm k_y y \pm k_z z)},$$

each of which can be associated with an electron moving in the direction of the vector with components $\pm k_x$, $\pm k_y$, $\pm k_z$.†

Thus the compounding, by *multiplication*, of three *independent* motions represented by uni-dimensional standing waves perpendicular to the coordinate axes is equivalent to the superposition, by *addition*, of eight *alternative* motions. All these eight motions correspond to the same absolute values of the momentum components g_x, g_y, g_z determined by the 'quantum conditions'

$$g_x = \pm \frac{h n_x}{2a}, \qquad g_y = \pm \frac{h n_y}{2b}, \qquad g_z = \pm \frac{h n_z}{2c}, \qquad (37\,a)$$

which follow from the 'nodal equations' (36 a). These quantum conditions are identical with the Bohr-Sommerfeld equations (32 a) for the particular case under consideration.

We have seen that in the uni-dimensional case each quantized state corresponds to a phase extension (i.e. an area in the classical phase plane x, g_x) equal to h. It can easily be seen that in the present three-dimensional case the phase extension (i.e. the volume of the six-dimensional space x, y, z, g_x, g_y, g_z) corresponding to one quantized state is equal to h^3. In fact equations (37 a) can be considered as defining a rectangular lattice of points in the three-dimensional momentum space

† It may be remarked that each pair of waves with opposite directions of propagation (for instance, k_x, k_y, k_z and $-k_x$, $-k_y$, $-k_z$) form together a uni-dimensional standing wave system with nodal planes dividing the edges a, b, c of the box into n_x, n_y, and n_z equal segments respectively.

g_x, g_y, g_z. The volume $\Delta g_x . \Delta g_y . \Delta g_z$ corresponding to each point in this momentum space is obviously equal to the volume of a parallelepiped with edges $h/2a$, $h/2b$, $h/2c$ forming the 'unit-cell' of the lattice. Multiplying this volume by the geometrical volume $a.b.c$ of the box we get for the phase volume corresponding to each point the expression $\Delta g_x . \Delta g_y . \Delta g_z . a.b.c = \frac{1}{8}h^3$. Now since each quantized state is represented by a system of eight symmetrical points (corresponding to the eight constituent running waves), the total phase extension corresponding to it must be equal to h^3.

The preceding results apply of course not only to electrons but to any other material particles, including photons. In the case of photons the associated waves can be considered as standing light waves in an empty box with perfectly reflecting walls—the 'Hohlraumstrahlung' (empty space radiation) of the German writers. Cathode and other material wave systems enclosed in a limited region of empty space can likewise be considered as normal modes of vibration, or 'characteristic' vibrations or 'eigenvibrations' (according to the German terminology) of this region, the nodal numbers characterizing these vibrations corresponding to the quantum numbers of the Bohr-Sommerfeld theory.

According to Schrödinger, similar 'eigenvibrations' in empty space can, under certain conditions, also take place in *infinite* space. The boundary represented by the walls of the box (or a vessel of any other shape) is not to be regarded as a geometrical surface, but as a dynamical surface-layer, on going through which the potential energy of the particle increases by a sufficiently large—strictly speaking infinite—amount. If now we imagine the thickness of the layer gradually increased until we obtain a stopping field of force continuously distributed throughout all space—as, for example, a Coulomb field of force about a positive nucleus—then the vibrations that we are considering *should not cease* but must merely alter their character gradually. The original bounding surface is thus displaced to infinity while the other nodal surfaces continuously alter their form. The *number* of nodal surfaces for any characteristic vibration must obviously remain constant. Thus we see that the number of different characteristic ('normal') vibrations must always, i.e. for any field of force pulling the particle inwards, be infinitely large.

The function ψ representing a characteristic vibration in three-dimensional space—whether limited or unlimited—can *always* be expressed as the product of three functions $\psi_1(q_1)$, $\psi_2(q_2)$, $\psi_3(q_3)$, corresponding to three *independent* components of the motion. Here q_1, q_2, q_3 are a set

of separable coordinates specifying these component motions, introduced before in connexion with the Bohr-Sommerfeld quantum conditions. The 'separability' of the coordinates q_1, q_2, q_3 is the mathematical expression of the independence of the corresponding component motions, the multiplicative combination of the functions $\psi_1(q_1)$, $\psi_2(q_2)$, $\psi_3(q_3)$ being in agreement with the multiplication law of the probability theory both in the old law for the probabilities, and in the new one for the probability amplitudes.

It must be remarked in conclusion that, so far as the combination of 'independent' motions is concerned, the theorem of the multiplication of probability amplitudes has been established without taking into account the time factor $e^{-i2\pi\nu t}$. This does not lead to errors in the evaluation of probabilities if only alternative motions with the *same energy* are considered. The multiplication theorem would be exactly valid if the energy ϵ—and therefore the frequency $\nu = \epsilon/h$—of a compounded motion was equal to the sum of the energies of the several independent motions constituting it.

Such an additive relation does hold in non-relativistic mechanics for the *kinetic* energy

$$W = \frac{1}{2m} g^2 = \frac{1}{2m} (g_x^2 + g_y^2 + g_z^2) = W_x + W_y + W_z,$$

where W_x, W_y, and W_z are those parts of W which correspond to the x, y, and z components of the motion. It does not hold, however, in relativistic mechanics where $\epsilon = c\sqrt{(m_0^2 c^2 + g^2)}$. In the limiting case of slow motion ($v \ll c$) ϵ can be represented as the sum of the additive kinetic energy $W = g^2/(2m)$ and the rest-energy $m_0 c^2$. Since the latter is the same for all states, the corresponding time factor $e^{-i2\pi m_0 c^2 t/h}$ can be dropped even when wave systems representing alternative motions with *different* energies are superimposed. Thus, neglecting relativity corrections, we can replace $\epsilon = m_0 c^2 + W$ in the relation $\nu = \epsilon/h$ by the additive energy W, without affecting the evaluation of probabilities. The multiplication theorem for motions in mutually perpendicular directions then becomes strictly valid. It must, however, be remembered that in principle, from the relativistic point of view, it is not strictly valid even in this case.

12. Reflection and Transmission Phenomena

In the preceding section we considered standing waves in a space limited by walls, playing the role of nodal surfaces and impermeable to the corresponding particles. Instead of this purely geometrical role we can

obviously ascribe to the walls a *dynamical* role, that of *reflecting* the travelling waves, by the superposition of which the standing waves are formed. This reflection is of course total, but of an idealized type which never occurs in nature, at least in the case of light waves. It can, indeed, be easily shown by the electromagnetic theory of light—as well as by the old mechanical theory—that in the case of total reflection the incident light waves penetrate into the second medium beyond the reflecting surface, their amplitude decreasing exponentially with increase of depth of penetration. Thus the reflecting surface does not actually coincide with the nodal surface which theoretically recedes to infinity (practically the vibrations die out at a distance comparable with the wave-length in the first medium).

This phenomenon can easily be understood if the reflecting surface is thought of as a plane, or rather as an infinitely thin layer with a *finite difference* of potential energy between its two sides. The idealized reflection described above corresponds to an *infinite* jump of potential energy, which would exclude the possibility of the transmission of waves for any angle of incidence. In the case of light this transmission does not take place if only the angle of incidence is larger than a limiting angle defined by the equation $\sin\alpha = \mu$. The index of refraction μ must of course be smaller than 1. It can easily be shown that from the corpuscular point of view total reflection actually takes place for angles of incidence for which the part of the kinetic energy corresponding to the normal component of the velocity of the incident light particle is smaller than the work to be done to pass through the surface, i.e. smaller than the increase of potential energy $U_2 - U_1 = U$ across it. We have in fact according to (2), § 1,

$$\mu = \sqrt{\left(1 - \frac{U}{\frac{1}{2}mv_1^2}\right)},$$

where v_1 is the velocity of the incident particle. The normal component of this velocity is equal to $v_1 \cos\alpha$. For the particle to be able to cross the surface we must have

$$\tfrac{1}{2}m(v_1\cos\alpha)^2 \geqslant U,$$

i.e.

$$1 - \sin^2\alpha \geqslant \frac{U}{\frac{1}{2}mv_1^2},$$

or

$$\sin\alpha \leqslant \sqrt{\left(1 - \frac{U}{\frac{1}{2}mv_1^2}\right)} = \mu,$$

the case $\sin\alpha > \mu$ corresponding to $\frac{1}{2}m(v_1\cos\alpha)^2 < U$.

According to classical mechanics the particle would be reflected when

$\sin\alpha > \mu$ without penetrating to the other side of the bounding surface, the value of the potential energy jump U being immaterial so long as the preceding inequality is fulfilled. This conclusion, however, contradicts the result obtained by treating the phenomenon of total reflection from the point of view of the wave theory, which, as mentioned above, requires a penetration of the vibrations beyond the surface. This penetration is practically limited to very small distances. Nevertheless, if we adopt the point of view of wave mechanics, that is, if we judge the behaviour of a particle by the results obtained for the associated waves, we must conclude that even when $\frac{1}{2}m(v_1\cos\alpha)^2 < U$ there is a certain probability of finding the particle beyond the reflecting surface, that is, in a region of space where the normal component of its velocity—or momentum—ought to *have an imaginary value*. This probability rapidly decreases with increasing distance from the surface.

This conclusion is of great importance, not only as an impressive example of the fundamental difference between the old and the new mechanics, but also because of its applications to various physical phenomena which were unintelligible from the old point of view. We shall therefore establish it in a more rigorous manner by developing quantitatively the theory of partial reflection and transmission, and of total reflection both of light and of matter waves.

For simplicity we shall limit ourselves to the one-dimensional problem, i.e. to the normal incidence of plane waves, travelling in the direction of the positive x-axis, on the plane $x = 0$, assuming the potential energy U_1 on the negative side of this plane to be zero, and on the positive side to be positive ($U_2 = U > 0$). If the kinetic energy of the incident particle $W = g^2/(2m)$ ($g = mv$ = momentum) is larger than U, then we shall have a partial reflection and a partial transmission of the associated waves, that is, a certain probability for reflection and another for transmission, the two probabilities adding up to unity. If $W < U$ (a case which can be realized in optics in the region of anomalous dispersion near an absorption line), we shall have total reflection, i.e. a probability 1 for reflection and a probability zero for transmission. Nevertheless we must be prepared to meet with a temporary penetration of the particle into the classically forbidden region.†

(*a*) Let us examine first the case $W > U$. The motion of the incident particle (instead of one particle we can imagine a stream of similar

† The reflecting surface behaves in this case like an elastic spring which, when hit by a moving particle, is compressed for a moment and then, as it recovers its original shape, throws the particle back.

particles all moving with the same velocity v_1 in the positive direction) is associated with a plane harmonic wave system represented by the function

$$\psi_1' = A_1' e^{i2\pi(k_1 x - \nu_1 t)}, \tag{38}$$

where the wave number $k_1 = 1/\lambda_1$ and the frequency are connected with the momentum and energy in the usual manner, i.e. by the 'quality' relations

$$hk_1 = mv_1, \qquad h\nu_1 = W = \tfrac{1}{2}mv_1^2.$$

In the latter relation we have dropped the irrelevant constant term $m_0 c^2$ (rest-energy), although its retention would not in the least complicate our calculations.

The wave function associated with the reflected particle (or the reflected part of the stream) must obviously have the form

$$\psi_1'' = A_1'' e^{i2\pi(-k_1 x - \nu_1 t)}, \tag{38 a}$$

and that associated with the transmitted particle (or the transmitted part of the stream) must have the form

$$\psi_2' = A_2' e^{i2\pi(k_2 x - \nu_2 t)}. \tag{38 b}$$

Here k_2—the wave number of the transmitted wave—is determined as usual by the relation

$$hk_2 = mv_2,$$

or, since

$$\tfrac{1}{2}mv_2^2 = W - U,$$

$$hk_2 = \sqrt{\{2m(W-U)\}}. \tag{39}$$

Obviously we must assume that the frequency ν_2 is equal to the frequency ν_1 of the incident waves (this is certainly so for light waves), and thus ν_2 is to be related not to the kinetic energy $W-U$ but to the total energy W, which remains constant, that is, has the same value on both sides of the reflecting surface. Thus

$$\nu_2 = \nu_1 = W/h. \tag{39 a}$$

k_1, k_2, and ν being known, our problem reduces to the determination of the amplitudes of the three waves or rather to the determination of the relative values of the amplitudes of the reflected and transmitted waves, the amplitude of the incident wave being taken as unity.

The ratio

$$R = \frac{|A_1''|^2}{|A_1'|^2} = \frac{|\psi_1''|^2}{|\psi_1'|^2} \tag{40}$$

which in terms of the wave theory can be defined as the reflection coefficient of the waves, in terms of the corpuscular theory obviously represents the probability that a particle striking the surface normally will be reflected from it. This probability must of course be less than 1. The difference

$$D = 1 - R \tag{40 a}$$

can be defined as the transmission coefficient of the waves or alternatively as the probability that the associated particle will be transmitted through the surface. It must clearly be proportional to the ratio $\frac{|A_2'|^2}{|A_1'|^2} = \frac{|\psi_2'|^2}{|\psi_1'|^2}$, but not equal to it, because the velocity of the particle has different values, v_1 and v_2, on different sides of the surface. The coefficient of proportionality can be determined by means of the following argument.

If instead of a single particle we think of a stream of particles and interpret accordingly the quantity $|\psi|^2 = \rho$ as the density of this stream, i.e. as the (relative) number of particles per unit volume, then the number of particles crossing unit area perpendicular to the direction of the stream per unit time will obviously be equal to the product of $|\psi|^2$ with the velocity of the particles v. This product

$$I = |\psi|^2 v \tag{41}$$

can be defined as the *density of the current* corresponding to the stream in question, or if we return to the initial interpretation of $|\psi|^2$ as referring to a single particle, then we can define it as the *density of the 'probability current'*.

Now clearly the transmission coefficient must be defined not as the ratio of the densities ρ_2 and ρ_1, but as the ratio of the corresponding current densities $\rho_2 v_2$ and $\rho_1 v_1$, i.e. as the ratio of the number of particles transmitted through the surface to the number striking it in the same time (the former being equal to the number of particles crossing any other surface parallel to the reflecting surface and lying beyond it). We thus get

$$D = \frac{I_2'}{I_1'} = \frac{|A_2'|^2 v_2}{|A_1'|^2 v_1}, \tag{41 a}$$

or, since v is proportional to the wave number k,

$$D = \left|\frac{A_2'}{A_1'}\right|^2 \frac{k_2}{k_1}. \tag{41 b}$$

The relation (40 a) can be interpreted as expressing the fact that, during a certain interval of time, the number of transmitted particles I_2' is equal to the difference between the number of incident particles I_1' and the number of reflected particles I_1''. In fact, since $R = I_1''/I_1'$, we have

$$\frac{I_2'}{I_1'} = 1 - \frac{I_1''}{I_1'} \quad \text{or} \quad I_2' = I_1' - I_1''.$$

This relation of course is not sufficient to determine the absolute values of the two ratios A_1''/A_1' and A_2'/A_1', to say nothing of the ratios themselves

(the amplitudes may in general be complex quantities). To determine these ratios we need *two linear* relations between the values of the three wave functions ψ_1', ψ_1'', and ψ_2', on both sides of the discontinuity formed by the reflecting surface $x = 0$. These relations cannot be deduced from the preceding considerations but must be introduced as a new postulate of the theory in such a way as to guarantee the validity of the quadratic relation (40 a).

The simplest and most natural relations of this kind are obtained by assuming that the resulting wave functions $\psi_1 = \psi_1' + \psi_1''$ and $\psi_2 = \psi_2'$ defining the motion on both sides of the discontinuity $x = 0$ can be considered as the values of one and the same wave function $\psi(x)$ defining the motion in all space and *remaining as continuous as possible at the boundary surface* $x = 0$. Since we can dispose of two relations only, these clearly must be the continuity condition for the function $\psi(x)$ itself, that is,

$$\psi_1 = \psi_2 \qquad (x = 0), \tag{42}$$

and for its first derivative, that is,

$$\frac{d\psi_1}{dx} = \frac{d\psi_2}{dx} \qquad (x = 0). \tag{42 a}$$

Putting $\psi_1 = \psi_1' + \psi_1''$ and $\psi_2 = \psi_2'$ we have, according to (38), (38 a), and (38 b),

$$A_1' + A_1'' = A_2',$$

$$(A_1' - A_1'')k_1 = A_2' k_2,$$

whence

$$A_1' = \tfrac{1}{2}\left(1 + \frac{k_2}{k_1}\right) A_2',$$

$$A_1'' = \tfrac{1}{2}\left(1 - \frac{k_2}{k_1}\right) A_2',$$

or

$$\left.\begin{aligned} \frac{A_1''}{A_1'} &= \frac{k_1 - k_2}{k_1 + k_2} \\ \frac{A_2'}{A_1'} &= \frac{2k_1}{k_1 + k_2} \end{aligned}\right\}. \tag{43}$$

This gives, according to the definitions (40) and (41 b),

$$R = \left|\frac{k_1 - k_2}{k_1 + k_2}\right|^2 \tag{43 a}$$

and

$$D = \frac{4k_1 k_2}{|k_1 + k_2|^2}. \tag{43 b}$$

Adding these two expressions we get $R + D = 1$ in agreement with (40 a). An interesting feature of the preceding expressions is their

symmetry with respect to k_1 and k_2. This symmetry shows that a particle moving with an original velocity $v_1 = hk_1/m$ from the negative side of the reflecting surface to the positive has the same probability of being reflected or transmitted (with a final velocity $v_2 = hk_2/m$) as a particle moving from the positive side to the negative with an original velocity v_2 (and a final velocity v_1 in case of transmission). This result corresponds to a well-known law of optics and represents, so far as transmission is concerned, the simplest case of a general law which will be referred to in future as the *law of reversibility* and which means that *transitions of opposite character have equal probabilities.*

It is interesting to notice that the partial reflection of the waves at the discontinuity of the potential energy is not due—as might have been expected—to the fact that the potential energy *increases* in the direction of motion, i.e. that the forces concentrated in the transition layer tend to *stop* the incident particle, for, as just pointed out, we get exactly the same value for the reflection coefficient for a motion in the opposite direction, when the forces in the transition layer accelerate the particle. To this we must add another remark concerning the experimental verification of the above conclusion in the case of cathode rays. The transition layer can be realized in this case by an electric condenser formed by two parallel plates of wire-gauze, so that the electrons could pass through the meshes. The two plates must of course be as near to one another as possible. Now according to our usual ideas, which seem to be well justified by experiment, no reflection of the cathode rays will take place unless they are completely stopped within the condenser, i.e. unless they have to pass across a potential energy jump U which is larger than their initial (kinetic) energy W. If $U < W$ (which is also the case when the potential difference in the condenser is not a stopping but an accelerating one), all the electrons should be transmitted—and experiment shows that they actually are.

This apparent discrepancy between the experimental facts and the wave-mechanical theory of the reflection and transmission of cathode rays through an electric condenser is due to the fact that we have been using a simplified model of such a condenser with the distance between the plates equal to zero. It will be shown later (Part II) that our conclusions remain approximately valid so long as the distance between the plates is small compared with the length $\lambda = h/mv$ of the cathode waves. In ordinary experiments of the above type we have to do, on the other hand, with macroscopic condensers, whose thickness is extremely large compared with λ. In this case wave mechanics leads to

practically the same results as classical mechanics. It is only in the realm of electric fields of microscopic extension such as are found in the surface-layer of metals and other material bodies, or in individual atoms and molecules, that the phenomena of reflection and transmission (or, as they are sometimes called, 'collision' phenomena) predicted by wave mechanics are really observed.

(*b*) Turning now to the case $W < U$, corresponding to total reflection, we are naturally led to assume that it can be dealt with by means of the same general equations as the case $W > U$. There is, however, one very important difference between the two cases; for the wave number k_2 as defined by equation (39) will now be an *imaginary* quantity:

$$k_2 = i|k_2|, \qquad |k_2| = \sqrt{\{2m(U-W)\}}. \tag{44}$$

This means, from the point of view of the wave theory, that the transmitted vibrations will be represented by a function

$$\psi_2' = A_2' e^{-2\pi|k_2|x} . e^{-i2\pi\nu t} \tag{44 a}$$

devoid of the usual oscillatory character in so far as its variation with x is concerned. Instead of a periodic variation we now have *an exponential decrease of the effective amplitude* $A_2' e^{-2\pi|k_2|x}$, in agreement with the description given at the beginning of this section.

Putting $k_2 = i|k_2|$ in the expression (43 a) for the reflection coefficient, since the conjugate complex of $k_1 \pm i|k_2|$ is equal to $k_1 \mp i|k_2|$, we get

$$R = 1,$$

which means that we now have total reflection. The amplitudes A_1' and A_1'' of the incident and the reflected wave have the same magnitude and differ from one another with respect to phase only.

Equation (43 b) gives an imaginary value for D which simply means that there is in the present case no transmission in the true sense of the word, but only a penetration with an exponentially decreasing amplitude. The 'effective depth' of this penetration can be defined as the reciprocal of the quantity $2\pi|k_2|$. It may be remarked that in the limiting case $U = \infty$, i.e. $|k_2| = \infty$, the depth of penetration reduces to zero, so that the reflecting surface becomes a nodal surface, as already assumed in the preceding section.

These results seem natural enough from the point of view of the wave theory but are rather paradoxical from the corpuscular point of view, since they imply that a particle can penetrate into regions of space where its kinetic energy, defined as the difference $W-U$, becomes negative, or in other words where its velocity, defined by the relation $v = hk/m$, becomes imaginary.

A possible way out of these difficulties, as well as the difficulties connected with the case $W > U$ (partial reflection instead of total transmission), lies in the restriction, left unnoticed so far, that we are dealing with wave systems defined for one-half of space only. Take, for instance, the incident wave system which is defined by equation (38) for negative values of x only. If we wish to extend our definition to positive values of x we must put

$$\left.\begin{aligned} \psi_1' &= A_1' e^{i2\pi(k_1 x - \nu t)} \qquad (x < 0) \\ \psi_1' &= 0 \qquad\qquad\qquad\quad (x > 0). \end{aligned}\right\} \tag{44 b}$$

Now it is clear that a wave system defined in this way cannot be considered as strictly harmonic—a harmonic system of plane waves being by definition unlimited. We can, however, represent the function (44 b) as a superposition of strictly harmonic waves with a continuously variable wave number k—in the same way as was done in the case of wave packets, or wave groups, limited on both sides (to a finite range Δx). This means that the velocity of the particles associated with the wave system (44 b), which is limited on one side, is actually not determined exactly, that is, $v_1 = hk_1/m$ is only the most probable value of the velocity. A similar argument can be applied to the reflected and the transmitted wave—if there is transmission. The important point is that the position of the incident, reflected, or transmitted particle is not completely undetermined (since it must be found *either* in the negative *or* in the positive half-space only), and so its velocity also cannot have an exactly determined value.† This might help us to understand why there is a partial reflection even in the case $W > U$: there is a certain probability that the incident particle will have a velocity insufficient to pass across the surface. The penetration of the particles beyond the reflecting surface in the opposite case (total reflection) can be explained in a similar way: no matter how large the jump of potential energy U may be, there is a non-vanishing probability that an incident particle will have sufficient kinetic energy to overcome it. The fading wave on the other side of the reflecting surface can also be represented by means of Fourier's integral as a superposition of harmonic waves associated with these transmitted particles.

The above considerations can be illustrated by the propagation of a wave packet or a *wave group* (of the type considered in § 9) in a field of force represented by our idealized condenser. However, we shall not discuss this question here, but shall merely remark that in applying the

† Or even *direction*.

above theory it is more convenient to adopt an unsophisticated attitude, taking the results at their face value and accepting the possibility of a negative kinetic energy or an imaginary velocity as a matter of course.

13. The Wave Equation of Schrödinger and its Application to the Harmonic Oscillator

The expressions (38), (38 a), (38 b) for the incident, reflected, and transmitted waves, as well as the boundary conditions (42) and (42 a), can be derived from a single differential equation for the resulting wave function ψ, i.e. the function $\psi = \psi_1' + \psi_1''$ for $x < 0$ and $\psi = \psi_2'$ for $x > 0$. This equation can be obtained by differentiating ψ *twice* with respect to x, which gives

$$\frac{d^2\psi}{dx^2} = -4\pi^2 k_1^2 \psi \qquad (x < 0)$$

$$\frac{d^2\psi}{dx^2} = -4\pi^2 k_2^2 \psi \qquad (x > 0)$$

(the second derivative being thus a discontinuous function of x at the point $x = 0$). Let us now replace k_1 and k_2 by their expressions

$$k_1 = \frac{mv_1}{h} = \frac{\sqrt{\{2m(W-U_1)\}}}{h}, \qquad k_2 = \frac{mv_2}{h} = \frac{\sqrt{\{2m(W-U_2)\}}}{h}, \tag{45}$$

where W—the total energy—is equal on both sides of the discontinuity $x = 0$ and is connected with the frequency by the relation†

$$\nu = \nu_1 = \nu_2 = W/h. \tag{45 a}$$

The function ψ can then be determined for all space by the differential equation

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2}[W - U(x)]\psi = 0, \tag{46}$$

with $U(x) = U_1$ for $x < 0$ and $U(x) = U_2$ for $x > 0$.

The continuity of the function ψ and its first derivative is automatically secured by the fact that the second derivative remains *finite* in virtue of equation (46) so long as U is finite throughout all space. Putting

$$\psi(x, t) = \psi^0(x) e^{-i2\pi\nu t}, \tag{46 a}$$

and taking into account the fact that ν is a constant independent of x, we can replace the wave function ψ in (46) by its amplitude $\psi^0(x)$ which is a function of x alone.

With this condition and with the function $U(x)$ of the form defined

† We are leaving out of account the 'rest-energy' $m_0 c^2$ which is an irrelevant constant, so far as the non-relativistic theory is concerned.

above, the general solution of equation (46) will be represented by the expressions (the superscript ^{0} will be dropped in future)

$$\psi_1(x) = A_1' e^{i2\pi k_1 x} + A_1'' e^{-i2\pi k_1 x} \qquad (x < 0),$$
$$\psi_2(x) = A_2' e^{i2\pi k_2 x} + A_2'' e^{-i2\pi k_2 x} \qquad (x > 0).$$

The second term in the latter expression represents a wave falling on the reflecting surface $x = 0$ from the positive side. The amplitudes A_2'' and A_1' can be chosen arbitrarily; the other two amplitudes will then be determined by the boundary conditions $\psi_1 = \psi_2$ and $\frac{d\psi_1}{dx} = \frac{d\psi_2}{dx}$ for $x = 0$. The resulting process can be represented as the superposition of two processes, one of which corresponds to $A_2'' = 0$ (waves falling from the negative side) and the other to $A_1' = 0$ (waves falling from the positive side).

Equation (46) was discovered by Schrödinger and bears his name. It is natural to assume that it can be applied not only to the simple type of potential energy function $U(x)$ which we have been considering so far, but also to a potential energy function of *any* form. It may be remarked that Schrödinger's equation is of the same type as the equation of the harmonic vibrations of a stretched string,

$$\frac{d^2\psi}{dx^2} + \frac{4\pi^2}{\lambda^2}\psi = 0, \tag{47}$$

where λ is the wave-length. In the simplest case of a uniform string (that usually considered in text-books) λ is constant, i.e. independent of x. The case studied above corresponds to two uniform strings of different density or thickness tied together at the point $x = 0$. Equation (47) remains, however, valid for a non-uniform string for which the wave-length λ (defined as the ratio of the frequency ν to the velocity of propagation at a given point) is any continuous or discontinuous function of x. In order to obtain Schrödinger's equation we need only take de Broglie's relation

$$\frac{1}{\lambda} = \frac{mv}{h}$$

between wave-length and velocity and express the velocity in terms of the potential energy $U(x)$. It must be noted, however, that in using this relation we actually go beyond the ideas originally implied in it, because we admit not only real but also *imaginary* values of the velocity corresponding to negative values of the kinetic energy $W - U$. In the case of non-relativistic mechanics, i.e. for (absolute) values of this

energy which are small compared with the 'rest energy' $m_0 c^2$, we can put

$$\frac{1}{\lambda} = \frac{\sqrt{\{2m(W-U)\}}}{h} \tag{47 a}$$

[cf. (45)]. If on the other hand we take the exact relativistic relation between momentum $g = mv$ and total energy† $\epsilon = m_0 c^2 + W = mc^2 + U$:

$$g^2 - \frac{(\epsilon - U)^2}{c^2} = -m_0^2 c^2,$$

we get

$$\frac{1}{\lambda^2} = \frac{1}{h^2 c^2}[(\epsilon - U)^2 - \epsilon_0^2], \tag{47 b}$$

where $\epsilon_0 = m_0 c^2$ is the rest energy (with $U = 0$). If $W - U$ is small compared with ϵ_0 we can put approximately

$$(\epsilon - U)^2 = (\epsilon_0 + W - U)^2 = \epsilon_0^2 + 2\epsilon_0(W - U),$$

in which case (47 b) reduces to (47 a).

The propagation of harmonic mechanical vibrations in a three-dimensional elastic medium is determined by a wave equation of the form

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} + \frac{4\pi^2}{\lambda^2}\psi = 0, \tag{48}$$

which is a generalization of the 'string-equation' (47). It is natural to assume that this 'wave equation' which is applicable both to mechanical and light vibrations should also be applicable to the vibrations associated, according to wave mechanics, with the motion of a particle in a given (constant) field of force. It must be emphasized that such a similarity between the mathematical equations determining the space-distribution of the three types of vibrations—sound, light, and matter waves—by no means implies a similarity between their natures or mechanisms. Nothing could be more erroneous than to mistake purely formal analogies for a genuine identity of physical nature. We must therefore not forget the difference between the physical meanings of the vibrating quantity $\psi = \psi^0(x, y, z)e^{-i2\pi\nu t}$ appearing in the 'wave equation' (48) in the case of sound, light, and matter waves. In the first case it means the displacement of a particle of the medium from its equilibrium position, in the second the electrical force at a given point (no mechanical motion being necessarily connected with it), and in the third the 'probability amplitude' of finding the particle associated with the waves at a given point. Of course the latter interpretation can be applied to light waves, leading to the idea of light quanta or photons associated with them. It could also be applied to sound waves,

† Including the potential energy U which so far has not appeared in ϵ.

in which case we should have to introduce the idea of 'sound quanta' or 'phonons' associated in a similar way with elastic vibrations. It seems of course very questionable whether such 'phonons' can have a real physical meaning. The same critical attitude should be preserved, however, with respect to light waves and perhaps to matter waves also —although for matter waves we seem to have direct evidence of the existence of the corresponding particles.

Taking this evidence for granted we can proceed to study the motion of such a particle in a given field of force specified by the potential energy $U(x, y, z)$, by investigating the solutions of equation (48) with λ determined according to (47 a) or (47 b). Denoting the differential operator $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$ by ∇^2 (∇ can be defined as the vector operator with components $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}$), we can write the resulting equations in the form

$$\nabla^2\psi+\frac{8\pi^2 m}{h^2}(W-U)\psi = 0 \tag{48 a}$$

(non-relativistic Schrödinger equation), and

$$\nabla^2\psi+\frac{4\pi^2}{h^2c^2}[(\epsilon-U)^2-\epsilon_0^2]\psi = 0 \tag{48 b}$$

(relativistic Schrödinger equation).

In order to check the Schrödinger equation and to illustrate the character of its solutions we shall apply it to two important special types of motion which are actually met with in nature, namely, the *harmonic oscillator* and the *hydrogen-like atom*.

A harmonic oscillator can be defined as a particle moving under a force which attracts it to a fixed point, the magnitude of the force being directly proportional to the distance of the particle from the point. If the motion is confined to one direction (or rather to two opposite directions), e.g. the x-axis, the oscillator is called *linear*. In this case the potential energy of the particle can be represented by

$$\tfrac{1}{2}\alpha x^2 \qquad (\alpha > 0), \tag{49}$$

the attracting centre being situated at the origin $x = 0$. If the motion can take place in any direction under a force corresponding to the potential energy

$$\tfrac{1}{2}\alpha r^2, \tag{49 a}$$

where $r^2 = x^2+y^2+z^2$, the oscillator is called a *spatial* oscillator, or more exactly an *isotropic* spatial oscillator. The more general case of

an *anisotropic* oscillator corresponds to a potential energy of the form

$$U = \tfrac{1}{2}(\alpha_1 x^2 + \alpha_2 y^2 + \alpha_3 z^2) \tag{49 b}$$

with three different (positive) coefficients α_1, α_2, α_3.

We shall consider first the linear oscillator. According to classical mechanics (neglecting relativity corrections) its motion is determined by the equation

$$m\frac{d^2x}{dt^2} = -\frac{\partial U}{\partial x} = -\alpha x$$

whose solution is $x = x_0 \cos(2\pi\nu t + \delta)$,

where the frequency ν is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{\alpha}{m}}. \tag{50}$$

The amplitude x_0, and consequently the energy

$$W = \tfrac{1}{2}m\left(\frac{dx}{dt}\right)^2 + \tfrac{1}{2}\alpha x^2 = \tfrac{1}{2}\alpha x_0^2 \tag{50 a}$$

is arbitrary.

According to wave mechanics the motion must be described by a wave function ψ, determined by the equation

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2}(W - \tfrac{1}{2}\alpha x^2)\psi = 0 \tag{51}$$

with the boundary conditions

$$\psi = 0 \qquad (x = \pm\infty). \tag{51 a}$$

These boundary conditions correspond to the boundary conditions $\psi = 0$ for $x = 0$ and $x = a$ which were introduced in the case, treated earlier, of a particle enclosed in a box with impermeable walls (§ 11), that is, for a potential energy function $U = 0$ for $0 < x < a$ and $U = +\infty$ for $x < 0$ and $x > a$. In the present case we have to do with a smooth function of the same type ($U = \infty$ for $x = \pm\infty$), the nodal planes being now displaced to infinity and the centre of motion being situated at the point (or plane) $x = 0$, instead of at $x = \frac{1}{2}a$.

For the sake of brevity we shall put

$$\frac{8\pi^2 m W}{h^2} = A, \qquad \frac{4\pi^2 m\alpha}{h^2} = B, \tag{52}$$

and so write the equation for ψ in the form

$$\frac{d^2\psi}{dx^2} + (A - Bx^2)\psi = 0. \tag{52 a}$$

To gain some idea of the nature of the solutions of this equation we shall first examine the behaviour of ψ for very large values of x, i.e. for

very large *negative* values of the kinetic energy $W-\frac{1}{2}\alpha x^2$. It will be remembered that these values correspond to the region of *total reflection* where, for a constant (finite) value of the potential energy, the amplitude of the vibrations ψ falls off exponentially with the distance. When x is large we may neglect A in comparison with Bx^2, and hence (52 a) reduces to

$$\frac{d^2\psi}{dx^2} = Bx^2\psi. \tag{52 b}$$

The 'asymptotic' form of the solutions of this equation as $x \to \infty$, and hence of equation (52 a), is

$$\psi = ce^{-\frac{1}{2}\sqrt{B}x^2}, \tag{53}$$

where c is an arbitrary constant. In fact, differentiating (53) we get

$$\frac{d^2\psi}{dx^2} = c(Bx^2-\sqrt{B})e^{-\frac{1}{2}\sqrt{B}x^2} = (Bx^2-\sqrt{B})\psi,$$

which is equation (52 a) with $\sqrt{B}$ instead of A, equivalent to (52 b) when x is large.

Having determined the asymptotic character of the function, i.e. its character for very large values of x, we can now proceed to determine it exactly by replacing the coefficient c in (53) by a suitably chosen function $c(x)$ of x. This function can be determined by substituting the expression (53) in the exact equation (52 a). Putting $B = \beta^2$ we have

$$\frac{d^2\psi}{dx^2} = c(\beta^2x^2-\beta)e^{-\frac{1}{2}\beta x^2} - 2\frac{dc}{dx}.\beta xe^{-\frac{1}{2}\beta x^2} + \frac{d^2c}{dx^2}e^{-\frac{1}{2}\beta x^2},$$

and consequently

$$\frac{d^2\psi}{dx^2} + (A-\beta^2x^2)\psi = e^{-\frac{1}{2}\beta x^2}\left[c(A-\beta) - 2\beta x\frac{dc}{dx} + \frac{d^2c}{dx^2}\right] = 0,$$

that is,

$$\frac{d^2c}{dx^2} - 2\beta x\frac{dc}{dx} + (A-\beta)c = 0. \tag{53 a}$$

This equation can be solved by expanding $c(x)$ in a power series

$$c = \sum_k a_k x^{\mu+k}, \tag{53 b}$$

where μ is a parameter to be determined later. Substituting (53 b) in (53 a) and equating to zero the coefficients of the various powers of x, we get the following 'recurrence relation' between the coefficients a_k:

$$a_{k+2} = \frac{2\beta(\mu+k)-(A-\beta)}{(\mu+k+2)(\mu+k+1)}a_k. \tag{54}$$

This relation shows that the series (53 b) only contains powers of x which differ from one another by even numbers, i.e. powers which are either odd or even, if μ is an integer. It shows further that this series

can be broken off both on the side of decreasing and on the side of increasing values of k, and so can be reduced to a *finite* number of terms ranging from $k_{\min}$ to $k_{\max}$. The limiting values of k are determined by the condition that a_k must vanish for $k = k_{\min}-2$ on the one hand and for $k = k_{\max}+2$ on the other, for this entails, according to (54), the vanishing of all the coefficients corresponding to smaller and to larger values of k. We thus get two 'boundary conditions' for the series (53 b):

$$(\mu+k_{\min})(\mu+k_{\min}-1) = 0 \tag{54 a}$$

and

$$A-\beta = 2\beta(\mu+k_{\max}),$$

i.e.

$$A = 2\beta(\mu+k_{\max}+\tfrac{1}{2}). \tag{54 b}$$

The former of these 'boundary conditions' is necessary to ensure a finite value of $\psi(x)$ for $x = 0$; it shows that we can put $\mu = 0$ and start the summation in (53 b) either with $k_{\min} = 0$ or $k_{\min} = 1$, so that the series (53 b) will contain positive powers of x only. The second 'boundary condition' (54 b) can be shown to represent the necessary consequence of the boundary condition (51 a) for the function ψ. In fact, if $c(x)$ does not reduce to a polynomial, that is, if the coefficient A does not assume one of the values (54 b), then *for sufficiently large values of* x the coefficients of the series (53 b) will satisfy approximately ('asymptotically') the relation

$$a_{k+2} \cong \frac{2\beta(\mu+k)}{(\mu+k+2)(\mu+k+1)}a_k \cong \frac{2\beta}{k+1}a_k,$$

which is satisfied also approximately by the coefficients a'_k in the expansion of the function

$$e^{+\beta x^2} = \sum_{s=0}^{\infty}\frac{(\beta x^2)^s}{s!} = \sum a'_k x^k \qquad (a'_{k+2} = \frac{\beta}{s+1}a'_k \cong \frac{2\beta}{k+1}a'_k,\ k = 2s).$$

From this it follows that, unless the series $c(x)$ terminates at some finite value of $k\,(= k_{\max})$, it must behave asymptotically for large values of x in a way similar to the function $e^{\beta x^2}$. The product $c(x)e^{-\frac{1}{2}\beta x^2} = \psi(x)$ would then, instead of tending to zero with $|x| = \infty$, on the contrary tend to infinity. Putting $k_{\max} = n$ we get, according to (54 b) and (52),

$$W = \frac{h}{2\pi}\sqrt{\left(\frac{\alpha}{m}\right)}(n+\tfrac{1}{2}),$$

that is,

$$W = W_n = h\nu(n+\tfrac{1}{2}). \tag{55}$$

Thus, unlike the classical theory, the wave-mechanical theory of the oscillator leads to discrete 'quantized' values of the energy just as for a particle moving in a box with perfectly reflecting walls (in the latter case the values of the energy are proportional to the square of the

quantum number n). In his original theory Planck obtained for the energy-levels of a harmonic oscillator the well-known formula $W = h\nu . n$ which differs from (55) by the absence of the term $\frac{1}{2}h\nu$. (The term $\frac{1}{2}h\nu$ corresponds to the 'second variant' of Planck's theory which implied the existence of a 'zero-point energy'.) Formula (55) was obtained by Heisenberg in his fundamental paper on quantum mechanics, and a little later by Schrödinger using a method equivalent to that described above.

The functions $c(x) = c_n(x)$ can be written in the form

$$c_n(x) = a_n^{(n)}x^n + a_{n-2}^{(n)}x^{n-2} + \ldots, \tag{55 a}$$

the last term being $a_1^{(n)}x$ when n is odd and $a_1^{(n)}$ when n is even. The coefficients $a_k^{(n)}$ are fixed in terms of one of them by the relation

$$a_{k+2}^{(n)} = -\frac{2\beta(n-k)}{(k+2)(k+1)}a_k^{(n)}, \tag{55 b}$$

which follows from (54) and (54 b). It can easily be shown that the polynomials (55 a) have n *real* roots lying symmetrically with respect to the point $x = 0$. Therefore the function $\psi_n^0(x) = c_n(x)e^{-\frac{1}{2}\beta x^2}$ can be represented by a wave-like curve, that is, a standing wave with $n+2$ nodal points, of which n are of the usual type and lie at a finite distance from and symmetrically with respect to the origin (if n is odd, one of these points coincides with the origin), and two lie at infinity ($x = \pm\infty$). The quantized or stationary states of a linear harmonic oscillator are thus associated with standing waves similar in type to those associated with the stationary states of a particle (electron) moving freely between two perfectly reflecting walls. The only difference between the two types of waves is that in the latter case the $n+2$ nodes are equidistant and extend over a finite length a, whereas in the present case their distance apart increases as we recede from the central point $x = 0$ and the two extreme nodal points corresponding to the walls lie at infinity. The polynomials $c_n(x)$ of the first five functions ψ_n^0 ($n = 0, 1, 2, 3, 4$) are respectively

$$c_0(x) = 1, \quad c_1(x) = 2\beta x, \quad c_2(x) = 4(\beta x)^2 - 2, \quad c_3(x) = 8(\beta x)^3 - 12\beta x,$$
$$c_4(x) = 16(\beta x)^4 - 48(\beta x)^2 + 12.$$

The first three of these functions are represented graphically in Fig. 4 (taken from Sommerfeld's book). The coefficient of proportionality is usually determined by the condition that the integral $\int_{-\infty}^{+\infty} |\psi_n(x)|^2\, dx$, which is a measure of the probability of finding the particle somewhere

on the x-axis between $+\infty$ and $-\infty$, in a state with a given energy W_n, should be equal to 1.

We shall now briefly examine the case of a spatial oscillator. In the classical theory this case differs from that of a linear oscillator only by

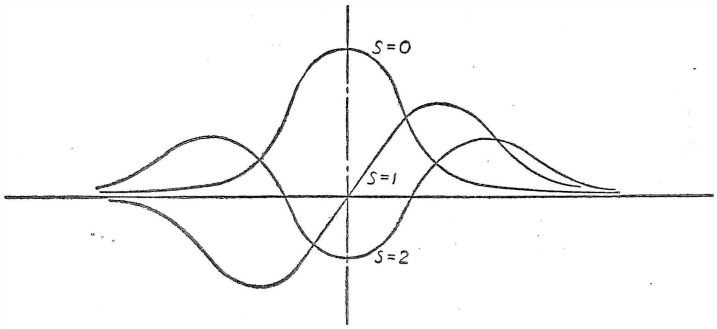

FIG. 4.

the fact that the general type of motion is not a rectilinear but an elliptic oscillation.

The wave-mechanical theory of a spatial oscillator reduces in the general case (anisotropic oscillator) to the solution of the equation

$$\nabla^2\psi+\frac{8\pi^2m}{h^2}(W-\tfrac{1}{2}\alpha_1x^2-\tfrac{1}{2}\alpha_2y^2-\tfrac{1}{2}\alpha_3z^2)\psi=0 \qquad (56)$$

with the boundary condition $\psi=0$ at infinity. Now if—as in the present case—the potential energy U, as well as the operator ∇^2, reduces to a sum of terms containing the separate coordinates only, the components of the motion with respect to these coordinates can be considered as *independent* uni-dimensional motions. This result (which can be generalized for any system of coordinates) is expressed in the language of the wave-mechanical theory by the '*multiplication law*' of the *probability amplitudes*, i.e. of the wave functions ψ_x, ψ_y, ψ_z, corresponding to the additive terms in the potential energy and the operator ∇^2. (We have considered this multiplicative relationship in § 11 for the special case $U=0$.†) In fact, putting

$$\psi(x,y,z)=\psi_x(x)\psi_y(y)\psi_z(z), \qquad (56\,a)$$

we get with the notation of (52)

$$\psi_y\psi_z\frac{\partial^2\psi_x}{\partial x^2}+\psi_z\psi_x\frac{\partial^2\psi_y}{\partial y^2}+\psi_x\psi_y\frac{\partial^2\psi_z}{\partial z^2}+(A-B_1x^2-B_2y^2-B_3z^2)\psi_x\psi_y\psi_z=0,$$

† The time factor $e^{i2\pi Wt/h}$ will be dropped in future. The multiplication law holds also for it, if the energy W is defined in the usual way without taking account of the 'rest energy'.

or, dividing through by ψ,

$$\left(\frac{1}{\psi_x}\frac{\partial^2\psi_x}{\partial x^2}-B_1x^2\right)+\left(\frac{1}{\psi_y}\frac{\partial^2\psi_y}{\partial y^2}-B_2y^2\right)+\left(\frac{1}{\psi_z}\frac{\partial^2\psi_z}{\partial z^2}-B_3z^2\right)+A=0.$$

Since x, y, and z are independent variables, it follows that each term in the brackets must be equal to a constant (otherwise it would depend upon the variables of the other terms which it does not contain). Denoting these constants by $-A_x$, $-A_y$, $-A_z$ we arrive at three equations similar to equation (52 a) for the linear oscillator, with the original A replaced respectively by A_x, A_y, A_z. The coefficient A is equal to the sum of A_x, A_y, A_z, corresponding to the additivity of the energy, i.e. $W=W_x+W_y+W_z$.

Writing $H_n(\beta x)$ for the polynomials $c_n(x)$ (since they actually depend upon the product βx only), we get

$$\psi_{n_x,n_y,n_z}(x,y,z)=H_{n_x}(\beta_1x)H_{n_y}(\beta_2y)H_{n_z}(\beta_3z)e^{-\frac{1}{2}(\beta_1x^2+\beta_2y^2+\beta_3z^2)}$$

and

$$W_{n_x,n_y,n_z}=h\nu_1(n_x+\tfrac{1}{2})+h\nu_2(n_y+\tfrac{1}{2})+h\nu_3(n_z+\tfrac{1}{2}),$$

ν_1, ν_2, ν_3 being the three fundamental frequencies for the oscillations in the directions of the coordinate axes.

In the special case of an isotropic oscillator we have $\beta_1=\beta_2=\beta_3=\beta$ so that the preceding formulae reduce to

$$\psi_{n_x,n_y,n_z}(x,y,z)=H_{n_x}(\beta x)H_{n_y}(\beta y)H_{n_z}(\beta z)e^{-\frac{1}{2}\beta r^2}, \qquad (56\,\text{b})$$

and $W=h\nu(n+\frac{1}{2})$ with $n=n_x+n_y+n_z+1$. This expression for the energy is identical with that corresponding to a linear oscillator.† However, for the linear oscillator different stationary states (characterized by the various values of the *single* quantum number n) have different energies, while in the present case the same value of the energy $W_n=h\nu(n+\frac{1}{2})$ corresponds in general to a number of different states, i.e. those for which $n_x+n_y+n_z=n-1$. This condition is termed *degeneracy*, and the states belonging to the same energy-level are termed degenerate. The number of degenerate states belonging to the same energy-level is called the *multiplicity* of the energy-level. To find the multiplicity of the level W_n, let us suppose that n_x has a fixed value. Then n_y can assume all values from 0 to $n-1-n_x$ (while n_z will vary in the opposite direction between the same limits). The number of states associated with a given value of n_x and n is thus equal to $n-n_x$. The total number of states associated with different values of

† For a 'plane' or two-dimensional oscillator we should obtain in the same way $W_n=h\nu n$ with $n=n_x+n_y+1$.

n_x from 0 to $n-1$, i.e. the multiplicity g_n of the energy-level W_n, is therefore equal to $\sum_{n_x=0}^{n-1}(n-n_x) = \sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$.

The existence of degeneracy is connected with an ambiguity in the specification of the stationary states. In the present case this ambiguity is represented by the arbitrariness of the choice of the coordinate axes. If we make a transformation to a new set of axes x', y', z' (with the same origin), the wave function (56 b) will be transformed to a sum of wave functions of the same type referring to the new coordinate axes with different values of the corresponding quantum numbers n'_x, n'_y, n'_z satisfying the condition $n'_{x'}+n'_{y'}+n'_{z'} = n_x+n_y+n_z$ and with appropriate numerical coefficients. In other words a 'given' degenerate state specified with respect to the old coordinate system by the quantum numbers n_x, n_y, n_z is equivalent to a set of alternative degenerate states specified with respect to the new coordinate system and belonging to the same energy-level W_n. This ambiguity can be traced even further if, for instance, we make a transformation from the original rectangular to polar coordinates. We shall consider this question in the next section in connexion with a similar problem for the hydrogen-like atom.

14. Wave Mechanics of a Hydrogen-like Atom; Comparison with Isotropic Oscillator and Rotator

A hydrogen-like atom can be defined as an electron moving in a Coulomb field of force attracting it towards a fixed centre, which can be visualized as a positive nucleus with a charge Ze (that of the electron being $-e$). The potential energy is then equal to

$$U = -\frac{Ze^2}{r}, \tag{57}$$

so that Schrödinger's equation assumes the form

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}\left(W + \frac{Ze^2}{r}\right)\psi = 0. \tag{57 a}$$

(We are dealing here, as before, with the non-relativistic equation.)

Since the potential energy does not reduce in this case to a sum of terms containing the separate Cartesian coordinates, it is more advantageous to use polar coordinates, for which this additivity condition is fulfilled (for the simple reason that U depends upon the radius r alone).

We shall first investigate the radially symmetrical solutions of equation (57 a), i.e. solutions of the form $\psi(r)$ which do not contain the angular variables defining the direction of the vector $\mathbf{r}$. It may be

remarked that the following analysis will, with certain modifications, also apply to the corresponding solution of the spatial oscillator problem if we replace the potential energy $-Ze^2/r$ by $\frac{1}{2}\alpha r^2$. There is, however, a fundamental difference between the two problems because for the oscillator the potential energy increases to infinity with increase of the distance r while for the hydrogen-like atom it tends to a *finite* limiting value, namely, $U = 0$. The 'elastic' field of force characterizing the oscillator is therefore similar to a box with perfectly reflecting walls in that it keeps the particle within a finite region for any value of the total energy W. The Coulombian field of the atom behaves in a similar way so long as W is smaller than $U_{\max}$, i.e. smaller than 0. If, however, $W > 0$, so that the electron can escape to infinity with a positive kinetic energy $W - U_{\max} = W$, it will do so, and the Coulombian field will produce an effect similar to the 'partial reflection' studied in § 12.

(*a*) In the special case under consideration (ψ function of r alone) the transformation to polar coordinates is particularly easy.

We have in fact

$$\frac{\partial\psi}{\partial x} = \frac{d\psi}{dr}\frac{\partial r}{\partial x} = \frac{d\psi}{dr}\frac{x}{r}, \qquad \frac{\partial^2\psi}{\partial x^2} = \frac{d^2\psi}{dr^2}\frac{x^2}{r^2} + \frac{d\psi}{dr}\left(\frac{1}{r} - \frac{x^2}{r^3}\right),$$

and consequently

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = \frac{d^2\psi}{dr^2} + \frac{2}{r}\frac{d\psi}{dr} = \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\psi}{dr}\right) = \frac{1}{r}\frac{d^2(r\psi)}{dr^2}. \tag{57 b}$$

If we put for brevity

$$r\psi = F, \qquad \frac{8\pi^2 m W}{h^2} = A, \qquad \frac{8\pi^2 m Z e^2}{h^2} = B, \tag{58}$$

then equation (57 a) can be written in the form

$$\frac{d^2F}{dr^2} + \left(A + \frac{B}{r}\right)F = 0. \tag{53 a}$$

When r is very large, i.e. at great distances from the nucleus, equation (58 a) practically reduces to the equation

$$\frac{d^2F}{dr^2} + AF = 0, \tag{59}$$

which determines the 'asymptotic' behaviour of the function F. Its general solution runs

$$F = c_1 e^{\sqrt{(-A)}r} + c_2 e^{-\sqrt{(-A)}r}. \tag{59 a}$$

According to our remarks above about the energy W we must distinguish between two cases, namely, $A < 0$ or $A > 0$. From the

corpuscular point of view the first case ($W < 0$) corresponds to a bound electron moving in an elliptical orbit, while the second corresponds to the hyperbolic motion of a free electron.

To begin with, we shall consider the first case and put

$$\sqrt{(-A)} = \alpha > 0.$$

Since the function ψ when $r \to \infty$ must remain finite, we must omit the first term in (59 a). We thus obtain

$$F = ce^{-\alpha r}, \tag{60}$$

that is, an exponential decrease of the amplitude of the waves (vibrations) in the region of 'total reflection'—in exact agreement with (44 a), § 12. The reason for such a quantitative agreement is that in both cases the potential energy can be considered as constant for large values of r (or x). We can represent in the form (60), not only the asymptotic solution of equation (58 a), but also its exact solution, if we take c to be a relatively slowly-varying function of r (for large values of r). We can determine this function by means of the equation

$$\frac{d^2c}{dr^2} - 2\alpha\frac{dc}{dr} + \frac{B}{r}c = 0, \tag{60 a}$$

which is obtained from (58 a) when we put $F = ce^{-\alpha r}$. The solution of this equation can be written as a power series

$$c = \sum_k b_k r^{\mu+k} \tag{60 b}$$

with an undetermined parameter μ which fixes the behaviour of the function $c(r)$ in the neighbourhood of the zero point.

By substituting this series in (60 a) and equating to zero the coefficients of different powers of r, we obtain the recurrence formula

$$b_{k+1} = \frac{2\alpha(k+\mu) - B}{(k+\mu)(k+\mu+1)} b_k. \tag{61}$$

In order that the function $\psi = ce^{-\alpha r}/r$ may remain finite throughout all space, the series (60 b) must break off at both ends, that is, it must reduce to a polynomial. If we put

$$c(r) = b_1 r^{\mu+1} + b_2 r^{\mu+2} + \ldots + b_n r^{\mu+n},$$

i.e. $b_0 = b_{n+1} = 0$, then from (61), because $b_1 \neq 0$ and $b_n \neq 0$, we get

$$\mu(\mu+1) = 0, \tag{61 a}$$

$$2\alpha(n+\mu) - B = 0. \tag{61 b}$$

Of the two solutions of (61 a) $\mu = 0$ is the only one consistent with

the condition of finiteness for ψ. Thus we obtain

$$c(r) = c_n(r) = b_1 r + b_2 r^2 + \ldots + b_n r^n, \tag{62}$$

and

$$\alpha \equiv \alpha_n = \frac{B}{2n}. \tag{62a}$$

The fact that the non-polynomial solutions of (60a) contradict the condition of finiteness for ψ, or more exactly the condition $\psi = 0$ for $r = \infty$, and hence do not come into consideration, can be seen as follows: If the condition (61b) is not fulfilled for any integral value of n, we obtain for $c(r)$ an infinite series with the asymptotic ratio of coefficients

$$\frac{b_{k+1}}{b_k} \simeq \frac{2\alpha}{k+1} \qquad \left(k \gg \frac{B}{2\alpha}\right).$$

It converges therefore in the same way as the series $\sum^{\infty} \frac{(2\alpha r)^k}{k!}$. Accordingly, when $r \to \infty$ the function $c(r)$ must behave like $e^{2\alpha r}$ and ψ must behave like $e^{+\alpha r}$. This is obviously not permissible.

We get, therefore, when $A < 0$, a discrete series of values of A given by the formula

$$A = -\alpha_n^2 = -\frac{B^2}{4n^2},$$

i.e., with reference to (58), by the formula

$$W_n = -\frac{2\pi^2 m Z^2 e^4}{n^2 h^2} \qquad (n = 1, 2, \ldots). \tag{63}$$

These 'characteristic' values of the energy agree exactly with those given by Bohr's theory if we regard n as the principal quantum number (defining the angular momentum $nh/2\pi$ in the case of circular orbits).

The corresponding functions $\psi_n(r) = F_n(r)/r$ can (on omitting the time factor) be written in the form

$$\psi_n(r) = \frac{e^{-\alpha_n r} c_n(r)}{r} = e^{-\frac{1}{2}\rho_n}(\gamma_1^{(n)} - \gamma_2^{(n)}\rho_n + \ldots + (-1)^{n-1}\gamma_n^{(n)}\rho_n^{n-1}) \tag{63a}$$

where

$$\rho_n = 2\alpha_n r, \qquad \gamma_{k+1}^{(n)} = \frac{n-k}{k(k+1)}\gamma_k^{(n)}. \tag{63b}$$

One of the coefficients $\gamma_k^{(n)}$ remains arbitrary. It can be fixed by the 'normalizing condition'

$$\int |\psi_n(r)|^2 \, dV = 4\pi \int_0^{\infty} |\psi_n(r)|^2 r^2 \, dr = 1, \tag{63c}$$

expressing the fact that the probability of finding the electron somewhere in the whole of space for a given stationary state is equal to 1.

It can be shown that the polynomial $c_n(r)$ has n *real* roots (including

$r = 0$). Therefore the wave function $\psi_n(r)$ represents *standing spherical waves* with $n-1$ *spherical nodal surfaces* at finite distances and a nodal surface at infinity. (Fig. 5 corresponds to the case $n = 4$.)

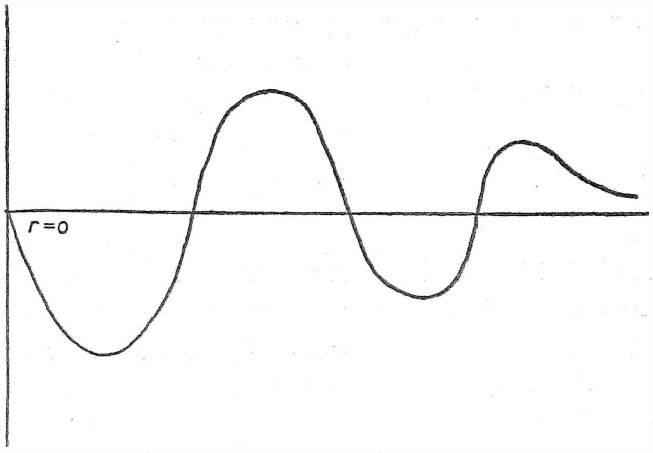

FIG. 5.

The normal state of the system is described by the (normalized) wave function

$$\psi_1(r) = \sqrt{\left(\frac{Z^3}{\pi a^3}\right)} e^{-Z\frac{r}{a}}, \tag{64}$$

where the constant

$$a = \frac{h^2}{4\pi^2 m e^2} \tag{64 a}$$

is Bohr's radius of the one-quantum orbit of the hydrogen atom ($Z = 1$). The energy of an elliptical orbit is expressed, according to classical theory, in terms of its semi-major axis a_n by means of the formula $W_n = -\frac{Ze^2}{2a_n}$. We have therefore, by (62 a) and (63 a),

$$a_n = \frac{n}{\alpha_n}. \tag{64 b}$$

As mentioned already, we can explain the exponential decrease of the functions $\psi_n(r)$ with increasing distance as the result of a total reflection of the spherical waves travelling outwards. The region where the wave-length—or in corpuscular mechanics the *velocity*—becomes imaginary begins at the distance $2a_n$. It is easy to see that the radius of the nodal surface lying farthest from the nucleus, i.e. the greatest root of the function $\psi_n(r)$, is of the same order of magnitude.

With positive values of the total energy W we obtain for ψ, by (59 a), the asymptotic expression

$$\psi = \frac{1}{r}(c_1 e^{i\sqrt{A}r} + c_2 e^{-i\sqrt{A}r}) \quad (A > 0). \tag{65}$$

If one of the coefficients c_1 and c_2 vanishes, then this expression represents spreading or converging spherical waves with a constant wavelength $\lambda = 2\pi/\sqrt{A}$. If we express A by means of (58) we obtain $\lambda = h/\sqrt{(2mW)}$. This relation agrees with the de Broglie formula $\lambda = h/(mv)$ if we notice that with $r = \infty$ the potential energy vanishes, so that W reduces to the kinetic energy $\frac{1}{2}mv^2$.

With $|c_1| = |c_2|$ we have standing waves extending to infinity. It can be shown that it is just this case which comes into consideration if ψ is supposed to represent a stationary state of the atom. In fact the quantities $|c_1|^2$ and $|c_2|^2$ measure the probability that an electron is travelling away from or approaching the nucleus respectively. In order that the probability of finding the electron at a given distance from the nucleus may remain constant, that is, independent of the time, the two opposite probability streams must be equal. If, for instance, $|c_1|^2$ were larger than $|c_2|^2$, then the integral $4\pi \int_0^r |\psi|^2 r^2\, dr$, measuring the probability of finding the electron somewhere within a sphere of radius r would decrease at a rate proportional to the 'current of probability' through this sphere, that is, at a rate proportional to the quantity

$$I = I_1 - I_2 = (|c_1|^2 - |c_2|^2)A$$

[cf. (41), § 12]. This would mean that $|\psi|^2$ would decrease exponentially with the time and therefore, since the dependence of ψ upon the time is given by the factor $e^{-i2\pi Wt/h}$, that the energy W would be a complex quantity (with a negative imaginary part). Apart from the fact that we are assuming W to be real, this would entail a complex value for A and, as can easily be shown, an exponential increase of ψ with increase of r, making $\psi = \infty$ for $r = \infty$. We shall consider this case at greater length in the next section in connexion with a simplified model of the hydrogen-like atom. For the present we shall merely notice that real values of W and finite values of ψ (vanishing at $r = \infty$) are consistent only with the condition $|c_1| = |c_2|$. This 'reality condition' must be satisfied not only asymptotically, i.e. at large distances, but for all values of r, if c_1 and c_2 are determined as functions of r by substituting (65) in the equation (57 a), or (59 a) in (58 a). In fact we get in this way

$$\left(\frac{d^2c_1}{dr^2} + 2i\sqrt{A}\,\frac{dc_1}{dr} + \frac{B}{r}c_1\right)e^{+i\sqrt{A}r} + \left(\frac{d^2c_2}{dr^2} - 2i\sqrt{A}\,\frac{dc_2}{dr} + \frac{B}{r}c_2\right)e^{-i\sqrt{A}r} = 0.$$

For this equation to hold for all values of r the two expressions in the brackets must vanish separately, so that we get for c_1 and c_2 two

separate equations of the type

$$\frac{d^2c}{dr^2} \pm 2i\sqrt{A}\frac{dc}{dr} + \frac{B}{r}c = 0. \tag{65 a}$$

It follows from these equations that if c_1 and c_2 have the same absolute value they must be *conjugate complex* quantities

$$c_1 = |c|e^{i\gamma}, \qquad c_2 = |c|e^{-i\gamma}. \tag{65 b}$$

The functions $c(r)$ can be expressed as before by *infinite* series, which, however, in contradistinction to the previous case ($A < 0$), do not reduce to polynomials and which with regard to their convergence character behave like the functions $e^{\mp 2i\sqrt{A}r}$. Since they are rather complicated we shall not consider them here. We shall merely record that ψ remains finite for finite values of r and vanishes at $r = \infty$ for *all* positive values of the energy W, so that the above condition for ψ does not entail a 'quantization' of the energy as in the case $W < A$. In addition, a normalization condition of the type used for the case of a bound electron, i.e. $\int |\psi|^2\, dV = 1$, is no longer possible since the integral $\int_0^r |\psi|^2 4\pi r^2\, dr$ becomes infinite as r tends to infinity unless the coefficients (or functions) c are assumed to be infinitely small. It should be remarked finally that the case $W > 0$, corresponding to a free electron or an ionized atom, lies beyond the scope of Bohr's quantum theory, which was restricted to motions for which the particle (or particles) remained within a limited region of space.

(*b*) The radially symmetrical solutions of the Schrödinger equation correspond, in classical mechanics, to *radial motions of the electron*, i.e. motions on straight lines which can be regarded as degenerate ellipses (if finite) or hyperbolas (if infinite), with vanishing minor axes. Bohr's theory excludes such straight-line orbits with a common origin in the nucleus. In the case of the bound electron it permits only elliptical orbits with a finite integral axis ratio k/n where the azimuthal quantum number k can assume values from 1 to n.

In wave mechanics, on the other hand, there is no reason at all for excluding the value $k = 0$. However, as we shall soon see, in wave mechanics the case $k = n$, which corresponds to a circular orbit, is excluded.

Corresponding to the non-radial motions of corpuscular mechanics—motions which are connected with a rotation—we must obviously, in wave mechanics, have functions ψ which depend not only on the distance r but also on the angles θ and ϕ which determine the direction

of the vector $\mathbf{r}$. If we define θ as the angle between $\mathbf{r}$ and the z-axis and ϕ as the angle between the plane $(\mathbf{r}, z)$ and the plane (x, z) (taken in the sense of a rotation from x to y), then we have

$$z = r\cos\theta, \qquad x = r\sin\theta\cos\phi, \qquad y = r\sin\theta\sin\phi \tag{66}$$
$$(x+iy = r\sin\theta\, e^{i\phi}),$$

and hence
$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2} = \frac{1}{r}\frac{\partial^2(r\psi)}{\partial r^2}+\frac{1}{r^2}\Omega^2\psi \tag{66 a}$$

with
$$\Omega^2\psi = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}. \tag{66 b}$$

The operator Ω^2 has therefore the same significance as the operator ∇^2 with regard to direction or 'angle' differentiation.

We thus obtain, instead of (58 a), the following differential equation for the function $F = r\psi$:

$$\frac{\partial^2 F}{\partial r^2}+\left(A+\frac{B}{r}\right)F+\frac{1}{r^2}\Omega^2 F = 0. \tag{67}$$

Its asymptotic solution for large values of r has the same form (59 a) as earlier with this difference, that the coefficients c_1 and c_2 are here certain functions of the angles θ, ϕ. Hence we seek to find the exact solution of (67) in the form $F = R(r)Y(\theta,\phi)$. If we substitute this expression in (67) and divide by F, then we get

$$\frac{r^2}{R}\left[\frac{d^2R}{dr^2}+\left(A+\frac{B}{r}\right)R\right] = -\frac{1}{Y}\Omega^2 Y.$$

Since the left side of this equation is independent of θ, ϕ and the right side is independent of r, both sides must be equal to one and the same constant Q. Thus equation (67) splits up into the following two equations:

$$\Omega^2 Y+QY = 0, \tag{67 a}$$

$$\frac{d^2R}{dr^2}+\left(A+\frac{B}{r}-\frac{Q}{r^2}\right)R = 0. \tag{67 b}$$

It can easily be seen that the first equation only possesses single-valued solutions when

$$Q = l(l+1), \tag{67 c}$$

where l is an arbitrary integer (which clearly need not be allowed to be negative).

To prove—or rather to illustrate—this assertion and at the same time to evaluate the corresponding functions $Y = Y_l(\theta,\phi)$ we return from the two-dimensional equation (67 a) to the three-dimensional equation (67) and consider the special case $A = B = 0$ [the constants

A, B are obviously immaterial for the functions Y and values of the constant Q in the 'angle equation' (67 a)].

In this case (67) reduces to the well-known Laplace equation $\nabla^2\psi = 0$ and the function ψ can be regarded (purely fictitiously) as the potential of a definite distribution of electricity at the point $r = 0$ or on the sphere $r = \infty$.

The simplest function of this kind is the potential of a point charge, namely, $1/r$. If we differentiate it with respect to one of the Cartesian coordinates x, y, z we obtain new solutions of Laplace's equation which correspond to the potentials of three elementary *dipoles* at the origin with moments parallel to the respective coordinate axes; the potentials are inversely proportional to r^2. Similarly, an arbitrary linear combination of the lth derivatives of $1/r$ with respect to x, y, z gives a further solution of the equation $\nabla^2\psi = 0$ of the form $\psi = \frac{1}{r^{l+1}} Y_l(\theta, \phi)$ which we can regard as the potential of an electric 'multipole of the lth order' (i.e. of a quadrupole when $l = 2$, an octupole when $l = 3$, etc.).

If we substitute this solution in the equation

$$\nabla^2\psi \equiv \frac{1}{r}\frac{\partial^2(r\psi)}{\partial r^2} + \frac{1}{r^2}\Omega^2\psi = 0$$

and take into account the relation

$$\frac{\partial^2(r\psi)}{\partial r^2} = \frac{l(l+1)}{r^{l+2}} Y_l(\theta, \phi),$$

we obtain the following equation for the 'direction factor' Y_l:

$$\Omega^2 Y_l + l(l+1)Y_l = 0.$$

When $Q = l(l+1)$ it completely agrees with (67 a). We may therefore identify the function $Y = Y_l$, which describes the dependence on direction of the Schrödinger wave function, with a so-called spherical harmonic of the lth order $Y_l(\theta, \phi)$ which occurs in the Coulomb potential of an electric multipole of the lth order.

We cannot develop here the general theory of spherical harmonics. This, however, is scarcely necessary if we remember their significance in the theory of the potential of electrical multipoles. For example, when $l = 1$, according to the well-known expression $\psi = \frac{p}{r^2}\cos(\mathbf{r}, \mathbf{p})$ for the potential of a dipole with electric moment $\mathbf{p}$ oriented in any direction, we have

$$Y_1(\theta, \phi) = \cos(\mathbf{r}, \mathbf{p}) = \frac{p_x}{p}\sin\theta\cos\phi + \frac{p_y}{p}\sin\theta\sin\phi + \frac{p_z}{p}\cos\theta.$$

Further, when $l = 2$,

$$Y_2(\theta, \phi) = \tfrac{1}{2}[3\cos(\mathbf{r}, \mathbf{p}_1)\cos(\mathbf{r}, \mathbf{p}_2) - \cos(\mathbf{p}_1, \mathbf{p}_2)],$$

where $\mathbf{p}_1$ and $\mathbf{p}_2$ are two arbitrary unit vectors which determine the axes of a quadrupole, and so on.

In the general case the function $Y_l(\theta, \phi)$ can be represented as the sum of $2l+1$ terms of the form

$$P_l(\cos\theta), \qquad \sin^m\theta\, P_l^{(m)}(\cos\theta)\cos m\phi, \qquad \sin^m\theta\, P_l^{(m)}(\cos\theta)\sin m\phi \tag{68}$$

with arbitrary coefficients c_l, A_l^m, B_l^m. Here $P_l(\cos\theta)$ is the so-called 'zonal' harmonic function or 'Legendre polynomial' of degree l, and $P_l^{(m)}(\cos\theta)$ is its mth derivative with respect to the variable $\cos\theta$.

The function $P_l(\cos\theta)$ describes the electric potential of a multipole with l-tuple z-axis. It can therefore be defined by the equation

$$P_l(\cos\theta) \sim r^{l+1}\frac{\partial^l(1/r)}{\partial z^l}, \tag{68 a}$$

where $z = r\cos\theta$. The other functions (68) can be derived most simply as those solutions of the equation $\Omega^2 Y_l + l(l+1)Y_l = 0$, which are equal to the product of $\cos m\phi$ or $\sin m\phi$ with some function of θ alone.

Instead of the real functions (68) it is possible, and indeed very natural in wave mechanics, to use complex functions:

$$\sin^m\theta\, P_l^{(m)}(\cos\theta)e^{im\phi}, \qquad \sin^m\theta\, P_l^{(m)}(\cos\theta)e^{-im\phi},$$

which can be multiplied with complex coefficients A_l^m, B_l^m. The number m can take all values between 0 and l. If negative values of m are used as well as positive in the interval

$$-l \leqslant m \leqslant +l,$$

then the above $2l+1$ harmonic functions can be represented by the single formula

$$Y_{lm} = \sin^{|m|}\theta\, P_l^{|m|}(\cos\theta)e^{im\phi}. \tag{68 b}$$

It should be noticed that the expression $\psi = \frac{1}{r}R(r)Y_{lm}$ holds both for $W < 0$ (bound electron) and $W > 0$ (free electron, ionized atom). These differ from one another only with respect to the radial function $R(r)$. When $W < 0$ we can put, as before,

$$R(r) = c(r)e^{-\alpha r} \qquad [\alpha = \sqrt{(-A)}]. \tag{69}$$

Substituting this expression in equation (67 b), and seeking $c(r)$ in the form of a power series $c = \sum b_k r^{\mu+k}$, we get for the coefficients b_k the recurrence formula

$$[(k+\mu+1)(k+\mu) - l(l+1)]b_{k+1} - [2\alpha(k+\mu) - B]b_k = 0. \tag{70}$$

This series can be reduced to an s-term polynomial

$$b_1 r^{\mu+1}+b_2 r^{\mu+2}+\ldots+b_s r^{\mu+s}$$

if we put $b_0 = 0$ and $b_{s+1} = 0$, i.e. if we put $(\mu+1)\mu-(l+1)l = 0$ on the one hand and $2\alpha(s+\mu)-B = 0$ on the other. This gives $\mu = l$ and

$$\alpha = \frac{B}{2(s+l)}. \tag{70a}$$

We thus obtain

$$c(r) = (2\alpha r)^l \sum_{k=1}^{s} (-1)^{k-1}\gamma_k^{(s)}(2\alpha r)^k \tag{71}$$

with $$\gamma_{k+1}^{(s)} = \frac{s-k}{(k+l)(k+l+1)-l(l+1)}\gamma_k^{(s)} \qquad (k \leqslant s). \tag{71a}$$

The sum $$s+l = n$$

plays the same role for the parameter α (and hence also for the energy) as the 'principal quantum number' introduced above. The numbers s and l correspond to the 'radial' and the 'angular' or 'azimuthal' quantum numbers of Bohr's theory. (The latter was previously denoted by k.) In Bohr's theory the smallest value of s was taken to be zero (circular orbit), and the smallest value of l was taken to be 1 (elliptical orbit with an axis ratio of $1/n$). In wave mechanics it is exactly the reverse. Here the smallest value of s is 1 [when the polynomial (71) is reduced to one term]. Corresponding to this the greatest permissible value of l is equal to $n-1$, while the smallest is zero ($s = n$). It will be remembered that in Bohr's theory the angular quantum number k served to define the angular momentum of the electron, which was supposed to be equal to $kh/2\pi$. The number m in (68b) represents, as we shall see later (Part II), the *axial* (or 'magnetic') quantum number of Bohr's theory: its product with $h/2\pi$ is considered there as the quantized value of the projection of the electron's angular momentum on the z-axis. The function ψ which corresponds to a given value of n and l is expressed by the formula

$$\psi \equiv \psi_{n,l,m} = e^{-\frac{1}{2}\rho_n}\rho_n^l\Big[\sum_{k=1}^{n-l}(-1)^{k-1}\gamma_k^{(n-l)}\rho_n^{k-1}\Big]Y_{lm}(\theta,\phi), \tag{72}$$

where $$\rho_n = 2\alpha_n r = \frac{8\pi^2 m Ze^2}{h^2 n} r = \frac{2n}{a_n} r. \tag{72a}$$

In the case $l = n-1$, which most nearly corresponds to the circular orbits of Bohr's theory, we have $\psi = \gamma_1^{(1)} e^{-\frac{1}{2}\rho_n}\rho_n^{n-1} Y_{n-1,m}(\theta,\phi)$. The maximum of the expression $|\psi|^2 4\pi r^2$ which multiplied by dr gives the probability for finding the electron at a distance between r and $r+dr$

from the nucleus (for a given direction θ, ϕ), is reached for a value of ρ_n defined by the equation $\frac{d}{d\rho}(e^{-\rho}\rho^{2n}) = 0$, which gives $\rho = 2n$, or according to (72a) $r = a_n$. We thus see that this 'most probable distance' for the state in question coincides with the radius of the corresponding circular orbit of Bohr's theory. The correspondence between states with lower values of l and Bohr's elliptical orbits is more complicated and will not be considered here.

We give below a list of the simplest functions ψ for values of $n < 4$. These functions† are normalized according to the condition

$$\int |\psi|^2 \, dV = 1.$$

For the sake of brevity ρ_n is written in the form $\rho_n = 2\sigma/n$, so that $\sigma = \alpha_1 r = Zr/a$ (a = radius of Bohr's one-quantum orbit in the hydrogen atom).

n l m	$\left(\frac{\pi a}{Z}\right)^{\frac{3}{2}} \psi_{n,l,m}$
1, 0, 0	$e^{-\sigma}$
2, 0, 0	$\frac{1}{4\sqrt{2}}(2-\sigma)e^{-\sigma/2}$
2, 1, 0	$\frac{1}{4\sqrt{2}}\sigma e^{-\sigma/2}\cos\theta$
2, 1, ± 1 . . .	$\frac{1}{8}\sigma e^{-\sigma/2}\sin\theta\, e^{\pm i\phi}$
3, 0, 0	$\frac{1}{81\sqrt{3}}(27-18\sigma+2\sigma^2)e^{-\sigma/3}$
3, 1, 0	$\frac{2}{81\sqrt{2}}\sigma(6-\sigma)e^{-\sigma/3}\cos\theta$
3, 1, ± 1 . . .	$\frac{1}{81}\sigma(6-\sigma)e^{-\sigma/3}\sin\theta\, e^{\pm i\phi}$
3, 2, 0	$\frac{1}{81\sqrt{6}}\sigma^2 e^{-\sigma/3}(3\cos^2\theta-1)$

The system of standing waves represented by $\psi_{n,l,m}$ has, besides the infinitely distant nodal surface $r = \infty$, also $s-1 = n-l-1$ spherical nodal surfaces, as well as l other nodal surfaces which are determined by the nature of the spherical harmonic $Y_l(\theta, \phi)$. In the simplest case they form $l-m$ cones θ = const. with the common axis z and, if real functions (68) are used, m meridian planes ϕ = const. which go through this axis. The energy W_n, or, in other words, the *vibration frequency* of the waves represented by the function $\psi_{n,l,m}$, depends only on the total

† Taken from Sommerfeld's *Ergänzungsband*, S. 84.

number of all the nodal surfaces and not on their distribution or form. We thus have a case of degeneracy similar to that considered in the preceding section in connexion with the isotropic spatial oscillator. The multiplicity of an energy-level W_n, i.e. the number of degenerate states belonging to it, is equal to the sum $\sum_{l=0}^{n-1}(2l+1)$, i.e. to n^2. The ambiguity of the degenerate states is partially expressed by the arbitrariness in the choice of the z-axis in reference to which the angular functions have been defined. If the direction of the polar axis is changed, a function $\psi_{n,l,m}$ representing a stationary state with respect to the original coordinate system is replaced by a sum of (in general n^2) functions $\psi_{n,l',m'}$ with suitable numerical coefficients $A_{l'}^{m'}$ representing stationary states belonging to the same energy-level in the new coordinate system. If the polar axis remains fixed but the two others (x, y) are rotated about it, $\psi_{n,l,m}$ is replaced by a linear combination of $2l+1$ functions $\psi_{n,l,m'}$. Similar results are obtained if we replace the polar coordinate system by some other system of 'separable' coordinates for which ψ can be represented as a product of functions of the separate coordinates (corresponding to independent component motions).† We shall not, however, pause to consider this question of the transformation of coordinates here.

As pointed out above, the representation of ψ in the form $\frac{R(r)Y_{lm}}{r}$ applies both to the case $W < 0$ and to the case $W > 0$. In the latter case the energy is not quantized, so that the principal quantum number n loses all meaning (unless it is defined as a transfinite number) and can be replaced by the continuously variable parameter W. It is, however, interesting to note that the other two quantum numbers, the 'angular' quantum number l and the 'axial' quantum number m, preserve their meaning and significance. In Bohr's theory the products of these numbers by $h/2\pi$ represent, as mentioned above, the angular momentum of the electron and its projection on the z-axis respectively. We can thus say that for the free electron or the ionized atom the angular momentum and its projection remain quantized in the same way as for the bound electron—with the important difference that the values of l and m are not limited. This means that the multiplicity of any positive energy-level is infinite—or, in other words, that the definition of a stationary state is 'infinitely ambiguous'.

† Such a system is provided by the so-called 'parabolic coordinates' defined by the equations $z = \frac{1}{2}(\lambda_1-\lambda_2)$, $x+iy = \sqrt{(\lambda_1\lambda_2)}e^{i\phi}$ [see Part III].

(c) It was mentioned, at the beginning of this section, that a bound electron in a Coulomb field of force is similar to an isotropic oscillator; in both cases the total energy is smaller than the maximum value of the potential energy and the latter is a function of the distance r alone. Hence it is clear that the wave functions of the isotropic spatial oscillator can also be represented in the form

$$\psi_{n,l,m} = \frac{1}{r} R(r)\, Y_{lm}(\theta, \phi),$$

where R will satisfy the equation

$$\frac{d^2R}{dr^2} + \left(A - Br^2 - \frac{Q}{r^2}\right) R = 0 \tag{73}$$

with $B = \dfrac{4\pi^2 m\alpha}{h^2}$, instead of (67 b), while the angular factor Y_{lm} has the same meaning as before. Putting $R = c(r)e^{-\frac{1}{2}\beta r^2}$ ($\beta = \sqrt{B}$) we get for $c(r)$ the equation

$$\frac{d^2c}{dr^2} - 2\beta r\frac{dc}{dr} + \left(A - \beta - \frac{Q}{r^2}\right) c = 0 \tag{73 a}$$

instead of (53 a), and the recurrence formula

$$a_{k+2} = \frac{\beta(\mu+k) - (A-\beta)}{(\mu+k+2)(\mu+k+1) - l(l+1)} a_k \tag{73 b}$$

instead of (54) [we have here replaced Q by its value $l(l+1)$]. This shows that in the series

$$c(r) = \sum a_k r^{\mu+k}$$

we must put $\mu = 0$, starting the summation from $k_{\min} = 0$ or 1. Denoting $k_{\max}$ by $n-l$ (instead of n as before) we get for $c_n(r)$ a polynomial of the nth degree with the lowest power of r equal to l or $l+1$. The corresponding energy value is determined by n alone in the same way as in the case of a uni-dimensional oscillator, i.e.

$$W_n = h\nu(n+\tfrac{1}{2}).$$

We thus see that the stationary states of a spatial oscillator can be specified by the quantum numbers n, l, m similar to those appearing in the theory of the hydrogen-like atom, i.e. quantum numbers defining the energy, the angular momentum and its projection on the z-axis respectively. The connexion of these quantum numbers with the quantum numbers n_x, n_y, n_z which have been used above is rather complicated with the exception of n, which is equal to $n_x+n_y+n_z+1$. The possibility of describing the stationary states of a spatial oscillator by functions of two entirely different types associated with different

types of independent component motions is again connected with the degeneracy, that is, with the fact that the energy-levels are not simple, but multiple. According to the preceding analysis, the multiplicity of a level W_n, that is, the number of degenerate states associated with it, was equal to n^2 for the hydrogen atom, but here rather surprisingly it has a different value $\frac{1}{2}n(n+1)$ in agreement with § 13. Whereas for the hydrogen atom l could take all values less than or equal to $n-1$, here the relation (73 b) is a relation between a_{k+2} and a_k, not a_{k+1} and a_k, and l can only take the series of integer values $n-1$, $n-3$, $n-5$,... (0 or 1). To each value of l there correspond the usual $2l+1$ different spherical harmonic functions and the total number of degenerate states is $\frac{1}{2}n(n+1)$, as we found before.

It should also be remarked that the preceding equation divided by r^2 can be considered as the wave-mechanical equation of motion for a particle capable of rotating about the origin of coordinates at a fixed distance $r = a$. Such a particle is called a (rigid) *rotator*. It can be fitted into the general scheme of Schrödinger's equation

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(W-U)\psi = 0$$

if we imagine that the potential energy U is equal to zero for $r = a$ and to infinity for $r \neq a$. In this case ψ must obviously vanish for $r \neq a$ and thus reduce practically to a function of the angles θ, ϕ alone. Dropping, then, the differentiation with respect to r in $\nabla^2\psi$, we get $\nabla^2\psi = \frac{1}{a^2}\Omega^2\psi$, i.e.

$$\Omega^2\psi + \frac{8\pi^2 m a^2}{h^2} W\psi = 0, \tag{74}$$

from which it follows that ψ must reduce (for $r = a$) to a spherical harmonic function $Y_l(\theta, \phi)$, giving for the (kinetic) energy the quantum condition

$$W_l = \frac{h^2}{8\pi^2 m a^2} l(l+1). \tag{74 a}$$

In Bohr's theory a rigid rotator was pictured as a particle rotating in a circle (one of the great circles of the sphere $r = a$) with a quantized angular momentum mva equal to $hl/2\pi$. Hence it followed that the (kinetic) energy of the rotator $W = \frac{1}{2}mv^2 = (mav)^2/2ma^2$ was equal to $h^2l^2/(8\pi^2 ma^2)$. The new wave-mechanical expression (74 a) differs from the old one by the fact that l^2 is replaced by $l(l+1)$. This difference corresponds to a fundamental difference in the treatment of motion in

classical and in wave mechanics. In the former case a 'spatial' rotator, i.e. a particle capable of assuming any position on a *sphere* of radius a, is actually treated as a *plane* rotator, i.e. a particle that can rotate along a definite circle of radius a (a great circle of the above sphere). If it were known that the particle was assigned to a particular circle, for example the equator $\theta = \frac{1}{2}\pi$ of the sphere, the operator Ω^2 would (for this circle) reduce to $\partial^2/\partial\phi^2$ and the wave equation to

$$\frac{\partial^2\psi}{\partial\phi^2} + \frac{8\pi^2 m W a^2}{h^2}\psi = 0.$$

Its solutions would then be of the type $\psi \sim e^{il\phi}$ with integral values of l to insure the single-valuedness of ψ (i.e. periodicity with respect to change of ϕ by an amount 2π). This gives

$$W = \frac{h^2 l^2}{8\pi^2 m a^2}$$

in agreement with the formula derived from Bohr's theory. The fact that the particle, representing the rotator, is not confined to a plane (circle), is expressed in the wave-mechanical theory by the appearance of a spherical harmonic $Y_{lm}(\theta, \phi)$ for the wave function instead of the circular harmonic $e^{il\phi}$. Putting $Y_{lm}(\theta, \phi) = \sin^{|m|}\theta P_l^{|m|}(\cos\theta)e^{im\phi}$ with m ranging from $-l$ to $+l$, we see that to each energy-level of the rotator specified by the angular quantum number l, there correspond $2l+1$ degenerate states specified by the various values of the axial number m, which determines the projection of the angular momentum on the z-axis. These $2l+1$ states represent various 'quantized' inclinations of the plane of the motion with respect to the z-axis ('spatial quantization' of the Bohr-Sommerfeld theory).

15. The Method of the 'Potential Staircase' for the Approximate Solution of one-dimensional Wave-Mechanical Problems

With the exception of a few simple problems of which the determination of the stationary states of the harmonic oscillator and of the hydrogen-like atom are the most important examples, the *exact* solution of Schrödinger's equation,

$$\nabla^2\psi + \frac{8\pi^2 m}{n^2}(W - U)\psi = 0$$

(with appropriate boundary conditions), presents insuperable mathematical difficulties. Different methods can be devised enabling one to find an approximate solution with any desired degree of accuracy. The simplest, though perhaps the crudest, of these methods is especially

convenient in the case of uni-dimensional problems when the potential energy U depends only upon *one* coordinate, x, say. It consists in replacing the actual continuous function $U(x)$ by a discontinuous function $\bar{U}(x)$, which has constant values within finite intervals and changes discontinuously from one of these values to another as we pass from one interval to the next. If U is represented graphically as a continuous curve $y = U(x)$, then $\bar{U}(x)$ will be represented by a curve like a staircase consisting of horizontal and vertical parts only, which follow more or less accurately the slope of the curve $U(x)$ (Fig. 6). This means that the continuous field of force defined by $U(x)$ is replaced by a series of force-free regions, divided from each other by infinitely thin partitions in which infinitely large forces are supposed to act in such a way that the work done by them is approximately equal to the work done by the actual force field in the interval Δx containing the corresponding partition. Such a method of approximation would be applicable both to the wave-mechanical and to the classical treatment of the motion, so long as in the classical treatment we were interested not in the acceleration of the particle but in its velocity; in the approximate field of force the velocity would not change continuously but would jump discontinuously from one constant value to another at the crossing of each partition. To this constant velocity there corresponds in the wave-mechanical treatment a constant wave-length. The wave function is represented in each region (s), for which $\bar{U}$ has the constant value $\bar{U}_s$, by an expression of the same type as if this region were infinite:

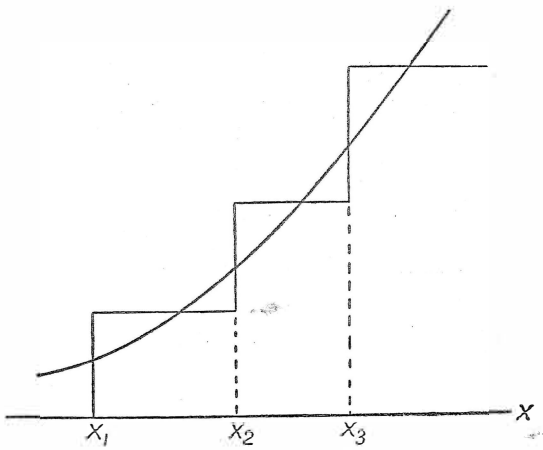

Fig. 6

$$\psi_s = (A'_s e^{i\alpha_s x} + A''_s e^{-i\alpha_s x}) e^{-i\omega t}, \tag{75}$$

with

$$\alpha_s^2 = \frac{8\pi^2 m}{h^2}(W - U_s), \qquad \omega = \frac{2\pi W}{h}. \tag{75a}$$

The constants A'_s, A''_s must be determined for the consecutive regions so as to ensure the continuity of the wave function and its derivative. Denoting by x_s the boundary between the sth and $(s+1)$th region,

we have
$$\left.\begin{aligned} A'_s e^{i\alpha_s x_s} + A''_s e^{-i\alpha_s x_s} &= A'_{s+1} e^{i\alpha_{s+1} x} + A''_{s+1} e^{-i\alpha_{s+1} x_s} \\ \alpha_s (A'_s e^{i\alpha_s x_s} - A''_s e^{-i\alpha_s x_s}) &= \alpha_{s+1} (A'_{s+1} e^{i\alpha_{s+1} x_s} - A''_{s+1} e^{-i\alpha_{s+1} x_s}) \end{aligned}\right\}. \quad (75\,\mathrm{b})$$

To these continuity conditions we must add the condition that ψ should remain finite at infinity:

$$\psi \text{ finite for } x = \pm\infty. \quad (75\,\mathrm{c})$$

The function ψ defined in this way, 'patched up' as it were of separate pieces, will be an approximation to the function corresponding to the continuous potential energy curve $U(x)$. By increasing the number of steps in the curve $\overline{U}(x)$ and decreasing the length $(x_{s+1} - x_s)$ and height $(U_{s+1} - U_s)$ of each step we can obtain any desired degree of approximation. All the fundamental properties of the solutions of Schrödinger's equation for a continuous function $U(x)$, in particular the existence (or non-existence) of a discrete set of 'characteristic' functions representing the quantized states of motion and belonging to discrete 'characteristic' values of the energy W, can easily be inferred from and interpreted by a study of the algebraic equations to which the determination of the approximate function is reduced.

Let us remark, to begin with, that the conditions (75 b) and (75 c) do not fix the absolute values of the coefficients A but only their *ratios* or *relative values*. (Conditions of this kind are called homogeneous.) This corresponds to the homogeneous linear character of Schrödinger's equation. Because of this character, if ψ is a solution representing some state of motion of the associated particle, the product of ψ by an arbitrary constant will also be a solution, *representing the same state of motion*. This fact (which has been particularly emphasized by Dirac) distinguishes fundamentally the wave-mechanical vibrations from the vibrations considered in the classical mechanics of wave motion or in the electromagnetic theory of light. We shall come back to this point later (Chapter IV).

The 'condition of finiteness' (75 c) restricts the coefficients A only if the limiting value of the potential energy at $x = -\infty$ or at $x = +\infty$, or both, is larger than the total energy W. This means that in the respective limiting region, extending from $x = -\infty$ to $x = x_1$ or from $x = x_{r+1}$ to $x = +\infty$ (the total number of regions being $r+2$), the parameter α becomes imaginary (total reflection) so that one of the two terms in (75) becomes infinite as x tends to $-\infty$ or $+\infty$. In virtue of (75 c) the coefficient of this term must be put equal to zero.

Let us suppose that this holds for both ends (which means that the

particle is in a 'bound' state, with an energy which is not sufficient to enable it to escape to infinity on either the positive or negative side). The number of unknown coefficients A is reduced from $2(r+2)$ to $2(r+2)-2 = 2(r+1)$. Now this is just equal to the number of boundary conditions (75 b). In view of their homogeneity the boundary conditions only determine the ratios of the coefficients A, so that actually the number of equations exceeds by 1 the number of unknown quantities. In order that the equations (75 b) should be compatible with each other, their determinant must vanish, and this condition gives a relation between the parameters α_1, α_2,..., α_{r+2}, that is, an equation for the energy W, in terms of which the α's are expressed by means of (75 a).

This equation allows, in general, an infinite number of discrete solutions, which represent the quantized energy-levels of the bound particle; to each of these energy-levels W_n there corresponds a definite system of coefficients A: A'_{ns}, A''_{ns}, etc., whose ratios are determined by the equations (75 b), or, in other words, a definite wave function $\bar{\psi}_n$ representing a stationary state of the particle. It should be noticed that this result holds independently of the shape of the curve $U(x)$ or $\overline{U}(x)$, assuming only that for some inner region U is smaller than at the two extremities ($x = \pm\infty$). It does not apply, however, when U has a local minimum which lies above the extreme values $U(\pm\infty)$, or above one of them. For then the kinetic energy $W-U(\pm\infty)$ will remain positive on both sides or at least on one, in which case the condition (75 c) does not restrict the coefficients A in the corresponding region. The number of unknown ratios between these coefficients thus becomes either equal to the number $2(r+1)$ of conditions (75 b) if $W-U > 0$ on one side and $W-U < 0$ on the other, or exceeds it by one, if $W-U > 0$ on both sides. In the first case the particle can be described as half-free (or half-bound), and in the second as completely free. An important feature of the above analysis is that it does not involve the intermediate portions of the curve $U(x)$; this means that the 'freedom' of the particle is not affected by the existence of local maxima of potential energy for which the kinetic energy $W-U$ is negative and through which it would be unable to pass according to classical mechanics. According to wave mechanics a particle is not bound when it is enclosed in a 'valley' of the curve $U(x)$—if the bottom of the valley lies above the levels of this curve at infinity. Let us suppose that $W < U(+\infty) = U^+$ and $W > U(-\infty) = U^-$ (half-free particle). In this case we get $A''_{r+2} = 0$, all the other coefficients A being determined by one of them and by the value of W which can be varied continuously so long as the above

conditions are fulfilled. The term $A_1' e^{i(\alpha_1 x - \omega t)}$ represents a wave, impinging, from an infinite distance on the negative side, on to the field of force represented by $U(x)$. This wave after a series of partial reflections is totally reflected at the other side of this field, so that the wave $A_1'' e^{i(-\alpha_1 x - \omega t)}$, which is the final result of this reflection and which is travelling back in the negative direction, must have the same amplitude as the incident wave and must form through superposition with the latter a *standing* wave system $A_1' \cos(\alpha x + \beta_1) e^{-i\omega t}$. The equality of amplitudes is expressed not by the equality of the coefficients A_1' and A_1'', which are complex quantities involving phase-constants, but by the equality of their moduli $|A_1'| = |A_2'|$. This result can be obtained directly from the equations (75 b) under the condition that α_{r+2} is imaginary and that $A_{r+2}'' = 0$. Further, it can easily be shown that the same relation holds not only for the first region but for all those regions in which the parameter α is real—in other words, that in all these regions we shall have standing waves of the type $A_s \cos(\alpha_s \omega t + \phi_s) e^{-i\omega t}$.

For a completely free particle the energy remains arbitrary, i.e. continuously variable, as in the preceding case, so long as the condition $W > U^{\pm}$ is fulfilled; the values of the $2(r+2)$ coefficients A are determined in this case not by one but by *two* of them, so that one of the ratios between these coefficients remains arbitrary. This arbitrariness can be removed by specifying the conditions under which the motion of the particle is supposed to take place.

We can suppose, *first*—as has been done in the case of the hydrogen atom with $W > 0$—that the motion represents a stationary state. In this case we must have, as before, standing waves in all the regions with a real parameter α. This result can be derived from an additional relation of the type

$$A_1'' = A_1' e^{i\gamma_1} \text{ or } |A_1''| = |A_1'|, \tag{76}$$

where γ_1 is an *arbitrary* (real) constant; this is equivalent to

$$\psi_1 = A_1 \cos(\alpha_1 x + \beta_1) e^{-i\omega t}$$

with an *arbitrary* phase-constant β_1. The corresponding phase-constants for the standing waves in the other regions will then be uniquely determined, as also will be the amplitudes A. For the inner regions where α is imaginary we shall have two real and, in general, different coefficients A' and A''.

Secondly, we can imagine that the motion represents the transmission or reflection of a particle coming into the field of force $U(x)$ from $x = -\infty$. In this case we can drop the term $A_{r+2}'' e^{-i(\alpha_{r+2} x + \omega t)}$ in the region extending to positive infinity, for this term represents a particle

coming into the field from the positive side. The additional condition

$$A''_{r+2} = 0 \tag{76a}$$

makes all the other coefficients expressible through one of them, A'_1, say; in particular the ratio A''_1/A'_1 will determine the reflection coefficient or the probability of reflection

$$R = \frac{|A''_1|^2}{|A'_1|^2} \tag{76b}$$

as in the simple case investigated in § 12, while

$$D = \frac{|A'_{r+2}|^2\alpha_{r+2}}{|A'_1|^2\alpha_1} = 1-R \tag{76c}$$

will give the probability of transmission. In the present case it is more convenient to contemplate a number of copies of the same particle instead of a single particle. The relative number of particles per unit volume associated with the wave system $A'e^{i(\alpha x-\omega t)}$ or $A''e^{-i(\alpha x+\omega t)}$ is then measured by $|A'|^2$ or $|A''|^2$.

It may be mentioned that the general solution of our problem when $W > U^{\pm}$ can be represented as a superposition of two solutions of the preceding type, corresponding to two oppositely directed incident streams of particles (or copies of a particle), i.e. with the condition $A''_{r+2} = 0$ for one stream and $A'_1 = 0$ for the other. If the two streams have the same intensity their superposition will represent a stationary state.

There exists still a *third* possibility for restricting the arbitrariness when $W > U^{\pm}$, to assume that at both ends there is an outgoing wave only, i.e.

$$A'_1 = 0, \qquad A''_{r+2} = 0. \tag{77}$$

This assumption is clearly inconsistent with the idea of a stationary state, for the number of particles (or copies of the given particle) which remain at any internal part of the field of force will obviously decrease as time goes on, if these particles stream outwards with no compensation in the shape of ingoing streams. Such a condition must be characterized from the wave-mechanical point of view by a decrease of the amplitude of the ψ-vibrations, that is, of the quantity $|\psi|^2$ with the time t. This can be obtained if we consider the vibrations as *damped*, i.e. define their frequency $\omega/2\pi$ as a *complex quantity*; from the corpuscular point of view this means that the *energy* $W = h\omega/2\pi$ must also be considered to be a complex quantity. After having been compelled to admit imaginary values of the velocity it does not seem unreasonable to admit complex values for the energy, which, by the way, according to relation

(75 a) and $v = h\alpha/(2\pi m)$, implies also complex values of the velocity. The physical meaning of these complex values is that they describe a state of affairs in which a number of similar particles, initially enclosed in a finite region of space limited by imperfectly reflecting walls, gradually *leak* through the walls and escape to infinity. The 'walls' can be pictured as local maxima, i.e. 'mountains' of potential energy, bounding a 'valley' whose bottom lies above the external flat parts ('plains') (Fig. 7).† Each particle is imagined to move in the appropriate

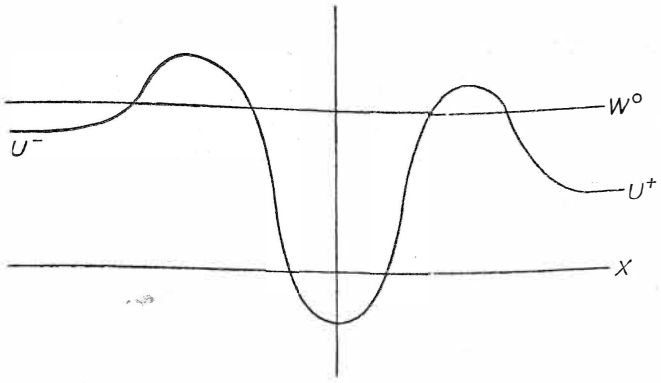

FIG. 7

field $U(x)$. All of these particles, together with their fields, are copies of one single system. The relative number of copies in a given state (with the particle inside or outside the valley, for example) simply measures the probability of this state for the system in question. Such a condition can be assumed to be realized when α-particles escape from the positive nuclei of radioactive atoms and cause in this way their disintegration (Gamow, Condon and Gurney). Out of a given number N of atoms (which can be considered as copies of one of them) a definite fraction, γN, say, disintegrate in unit time. The fact that this fraction does not depend upon the mutual action of the atoms nor on any external conditions must be interpreted as the expression of an inherent tendency of each atom to disintegrate. From this point of view the constant γ—the so-called *decay constant*—must be associated with an individual atom and interpreted as the probability of its disintegration per unit of time. Such a state of affairs was very puzzling for classical mechanics (which would require the atom either to disintegrate instantaneously or not to disintegrate at all), but it fits in very well with the general conceptions of wave mechanics. It can be explained quantitatively by

† This means that the particle is attracted to the centre in the internal and repelled from it in the external region.

assuming the field of force exerted by the nucleus on the α-particle to be of the type of Fig. 7, and by ascribing to the α-particle a complex energy

$$W = W^0 - iW' \tag{77 a}$$

with $W^0 > U^{\pm}$ and a small positive value of W'. According to the formula $\psi = \psi^0(x)e^{-i2\pi Wt/h}$ this gives for the integral $\int |\psi|^2\, dx$ taken over the internal region of the field ('valley') a formula containing a factor $e^{-\gamma t}$ with

$$\gamma = \frac{4\pi W'}{h}. \tag{77 b}$$

This integral measures the relative number of particles which can be found at time t within the respective nuclei (i.e. the number of atoms that have not suffered a disintegration). The quantity W', and consequently the decay (or 'leaking') constant γ, can be determined by the approximate method which we are considering in this section from the boundary conditions (75 b) and the additional conditions (77), which provide, just as for a completely bound particle, an equation for the energy W. This equation, in view of the complex value of W, splits up into two equations which serve to determine both the real and the imaginary part of W. We thus get a relation between the decay constant γ of different radioactive atoms and the kinetic energy of the escaping particles W^0 of the same type as that empirically established by Geiger and Nuttall. The above method (due to Gamow) is open to some objections. First of all, as has been mentioned above, the multiplication of ψ by a factor independent of x—whether this factor depends on the time or not—does not change the state represented by ψ. Further, the introduction of complex energy-values necessitates the introduction of complex values for the parameters α, in particular for the parameters α_1 and α_{r+2} characterizing the two outgoing waves. These complex values involve real exponential factors in the expressions

$$A_1'' e^{-i\alpha_1 x}, \qquad A_{r+2}' e^{i\alpha_{r+2} x},$$

characterizing the outgoing waves, which make ψ infinite both at $x = +\infty$ and $x = -\infty$. This result, which is a violation of the finiteness condition (75 c), follows without any calculation from the fact that the particles to be found at large (infinite) distances are those that escaped from the central part of the field a long (infinite) time ago when they were practically all within it.

Just as in the simple case of reflection and transmission due to a 'potential energy jump', investigated in § 12, we can get rid of complex energies and velocities by representing the separate 'pieces' of the func-

tion ψ limited to the regions for which the potential energy is assumed to be constant as a superposition of unlimited plane waves with different *real* values of the wave numbers $\alpha/2\pi$ and frequencies $\omega/2\pi$.

Very instructive results are also obtained by considering the motion, in a field of the type shown in Fig. 7, of a wave group initially confined to the central region. We shall not, however, investigate these questions, but shall give a few simple examples of the application of the method described in this section to fields of various kinds.

16. Application of the Preceding Method to some Simple Problems

1. *Simplified Uni-dimensional Model of a Hydrogen-like Atom.* The potential energy $U(x)$ will be represented by the following function (Fig. 8):

$$U = -V \text{ for } |x| < l; \qquad U = 0 \text{ for } |x| > l.$$

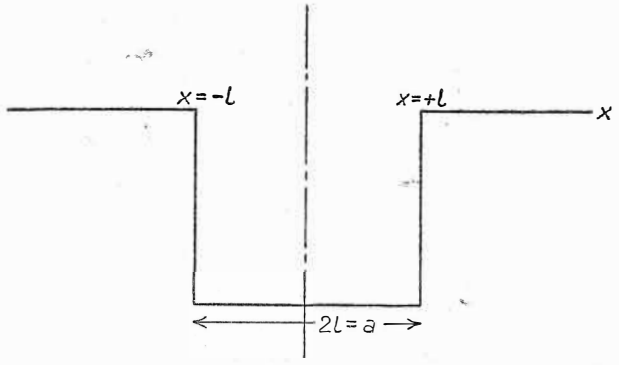

FIG. 8

Case (a): $W < 0$. We have for the internal region ($s = 0$) (dropping the time factor $e^{-i\omega t}$)

$$\psi_0 = A_0 \cos(\alpha_0 x + \phi_0),$$

and for the external regions ($s = -1, +1$)

$$\psi_{\pm 1} = A_{\pm 1} e^{\mp |\alpha_1| x},$$

where

$$\alpha_0 = \sqrt{\left\{\frac{8\pi^2 m}{h^2}(W+V)\right\}}, \qquad |\alpha_1| = \sqrt{\left\{-\frac{8\pi^2 m W}{h^2}\right\}} = \beta_1.$$

The upper sign refers to $x \geqslant l$, and the lower to $x \leqslant -l$. The boundary conditions (75 b) reduce to the following equations:

$$A_0 \cos(\alpha_0 l + \phi_0) = A_{+1} e^{-\beta_1 l},$$

$$\alpha_0 A_0 \sin(\alpha_0 l + \phi_0) = \beta_1 A_{+1} e^{-\beta_1 l},$$

for $x = +l$, and

$$A_0 \cos(-\alpha_0 l + \phi_0) = A_{-1} e^{-\beta_1 l},$$
$$-\alpha_0 A_0 \sin(-\alpha_0 l + \phi_0) = \beta_1 A_{-1} e^{-\beta_1 l},$$

for $x = -l$. Dividing the second equation of each group by the first, we get

$$\tan(\alpha_0 l + \phi_0) = \frac{\beta_1}{\alpha_0} = -\tan(-\alpha_0 l + \phi_0). \tag{78}$$

Hence it follows that ϕ_0 is equal either to 0 or to $\frac{1}{2}\pi$, so that $\psi_0 = A_0 \cos \alpha_0 x$ (standing waves symmetric with respect to the origin) or $\psi_0 = A_0 \sin \alpha_0 x$ (waves antisymmetric with respect to the origin). In the first case we have

$$A_{+1} = A_{-1} = e^{\beta_1 l} A_0 \cos \alpha_0 l,$$

while in the second $A_{-1} = -A_{+1}$. The quantized energy values are given (in conjunction with the definitions of α_0 and β_1) by the equation (78) which reduces to

$$\tan \alpha_0 l = \frac{\beta_1}{\alpha_0} \tag{78 a}$$

in the symmetrical, and to

$$\cot \alpha_0 l = -\frac{\beta_1}{\alpha_0} \tag{78 b}$$

in the antisymmetrical case. These equations have a limited number of solutions so long as V has a finite value. They can be obtained graphically by plotting the curves $\tan \xi$ or $-\cot \xi$ and finding their points of intersection with the curve $\sqrt{(a^2 - \xi^2)}/\xi$ where $a^2 = 8\pi^2 m V l^2/h^2$. The quantized values of the energy are given by

$$W = \frac{-n^2(a^2 - \xi^2)}{8\pi^2 m l^2}.$$

The largest of them cannot of course exceed 0, and the smallest cannot be less than $-V$. In the limiting case $V = \infty$ we get $\tan \xi = \infty$, $\xi = \frac{1}{2}(2n' + 1)\pi$, and therefore $\lambda_0 = \frac{2\pi}{\alpha_0} = \frac{4l}{2n' + 1}$ for the symmetrical solutions, and similarly $\cot \xi = \infty$ for the antisymmetrical solutions. This agrees with the results found in § 11 for standing waves in a box with perfectly reflecting walls, if $2l$ is the length of the box, and $n = 2n' + 1$ or $2n''$ the quantum number, i.e. the number of internal nodes. The actual functions $\psi(x)$ corresponding to the approximate functions (for $V = \infty$) with $n = 2$ and $n = 3$ are shown in Fig. 9. It will be noticed that the curves pass continuously and with a continuous tangent from the central region into the outward ones.

Case (*b*): $W > 0$. The stationary states are described by standing

waves so that $\psi_s = A_s \cos(\alpha_s x + \phi_s)$, where $s = -1, 0, +1$ and $\alpha_{-1} = \alpha_{+1}$. One of the phase-constants ϕ_s can be chosen arbitrarily. If we confine ourselves to symmetric and antisymmetric solutions (any other solution can be represented as a superposition of these two), i.e. to $\phi_0 = 0$ and $\phi_0 = \frac{1}{2}\pi$, we get from the boundary conditions $\psi_s = \psi_{s+1}$ and $d\psi_s/dx = d\psi_{s+1}/dx$ for $x = +l$ and $\phi_0 = 0$:

$$A_0 \cos \alpha_0 l = A_{+1} \cos(\alpha_1 l + \phi_1),$$
$$\alpha_0 A_0 \sin \alpha_0 l = \alpha_1 A_{+1} \sin(\alpha_1 l + \phi_1),$$

whence

$$\tan \alpha_0 l = \frac{\alpha_1}{\alpha_0} \tan(\alpha_1 l + \phi_1),$$

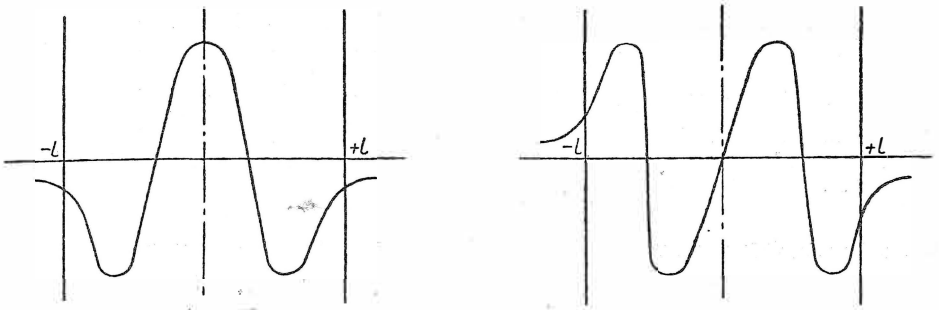

FIG. 9

and similarly, from the boundary conditions at $x = -l$,

$$\tan \alpha_0 l = \frac{\alpha_1}{\alpha_0} \tan(\alpha_1 l - \phi_{-1}).$$

These equations do not restrict the values of the parameters α, and consequently of the energy W, but merely serve to determine the phase-constants $\phi_{\pm 1}$. It is obvious from them that for the symmetrical case, $\phi_{-1} = -\phi_{+1}$, and $A_{-1} = A_{+1}$. In the same way we find for the antisymmetrical case $\phi_{-1} = -\phi_{+1}$ and $A_{-1} = -A_{+1}$.

We shall now examine the motion specified by the subsidiary condition $A''_{+1} = 0$. We thus consider a particle travelling from left to right into the potential valley (representing the 'interior' of the atom) and either emerging from it on the other side or being thrown back (reflected). Our problem will now be to calculate the probability of reflection R or of transmission D. The boundary conditions (75 b) reduce in the present case to

$$A'_0 e^{i\alpha_0 l} + A''_0 e^{-i\alpha_0 l} = A_{+1} e^{i\alpha_1 l}$$
$$\alpha_0 (A'_0 e^{i\alpha_0 l} - A''_0 e^{-i\alpha_0 l}) = \alpha_1 A_{+1} e^{i\alpha_1 l}$$

for $x = l$, and

$$A'_{-1}e^{-i\alpha_1 l}+A''_{-1}e^{+i\alpha_1 l} = A'_0 e^{-i\alpha_0 l}+A''_0 e^{+i\alpha_0 l}$$
$$\alpha_1(A'_{-1}e^{-i\alpha_1 l}-A'_{-1}e^{+i\alpha_1 l}) = \alpha_0(A'_0 e^{-i\alpha_0 l}-A''_0 e^{i\alpha_0 l})$$

for $x = -l$.

It will be convenient to choose the coefficient A_1 $(= A'_1)$, representing the amplitude of the transmitted wave, as the known quantity, and express the other coefficients in terms of it. Actually we do not need the coefficients A'_0, A''_0, which must be eliminated by solving the first two equations and substituting the resulting expressions in the second two. We thus get

$$A'_0 = \tfrac{1}{2}A_{+1}\left(1+\frac{\alpha_1}{\alpha_0}\right)e^{i(\alpha_1-\alpha_0)l}, \qquad A''_0 = \tfrac{1}{2}A_{+1}\left(1-\frac{\alpha_1}{\alpha_0}\right)e^{i(\alpha_1+\alpha_0)l},$$

and then

$$A'_{-1} = \tfrac{1}{4}A_{+1}\left[\left(1+\frac{\alpha_1}{\alpha_0}\right)\left(1+\frac{\alpha_0}{\alpha_1}\right)e^{i2(\alpha_1-\alpha_0)l}+\left(1-\frac{\alpha_1}{\alpha_0}\right)\left(1-\frac{\alpha_0}{\alpha_1}\right)e^{i2(\alpha_1+\alpha_0)l}\right],$$

$$A''_{-1} = \tfrac{1}{4}A_{+1}\left[\left(1+\frac{\alpha_1}{\alpha_0}\right)\left(1-\frac{\alpha_0}{\alpha_1}\right)e^{-i2\alpha_0 l}+\left(1-\frac{\alpha_1}{\alpha_0}\right)\left(1+\frac{\alpha_0}{\alpha_1}\right)e^{i2\alpha_0 l}\right],$$

or writing $2l = a$ for the width of the valley,

$$A'_{-1} = A_{+1}\left[\cos\alpha_0 a-\tfrac{i}{2}\left(\frac{\alpha_1}{\alpha_0}+\frac{\alpha_0}{\alpha_1}\right)\sin\alpha_0 a\right]e^{i\alpha_1 a}, \tag{79}$$

$$A''_{-1} = \tfrac{1}{2}A_{+1}\left[\frac{\alpha_1}{\alpha_0}-\frac{\alpha_0}{\alpha_1}\right]\cos\alpha_0 a.$$

The probabilities R and D can be calculated from these expressions by means of (76 b) and (76 c). The interesting point about them is that R vanishes, i.e. we get complete transmission, when the energy W of the incident particle is such that $\cos\alpha_0 a = 0$, that is, when $\alpha_0 a = \frac{1}{2}(2n+1)\pi$, where n is an arbitrary integer. This means that the wave-length of the particle within the depression $\lambda = 2\pi/\alpha_0$ corresponding to complete transmission must be equal to $2a/(2n+1)$. In other words, it must coincide with the wave-length of one of the symmetrical standing waves, representing the stationary states of the particle when $V = \infty$. The 'quantized' values of the energy W (which can be identified with the kinetic energy of the particle outside the U-valley) with respect to this 'selective' transmission are given by the equation

$$\frac{8\pi^2 m}{h^2}(W+V) = \frac{(2n+1)^2\pi^2}{4a^2},$$

which follows from (75 a) in conjunction with $\alpha_0 a = \frac{1}{2}(2n+1)\pi$, or

$$W_n = \frac{h^2}{8ma^2}(n+\tfrac{1}{2})^2 - V. \tag{79 a}$$

The number n must of course be large enough to make W_n positive. This means that it must be larger than for any of the (approximately defined) stationary states corresponding to $W < 0$ (bound particle).

The preceding results can easily be generalized to the case when the potential energy has different values on the two sides of the valley (i.e. when $\alpha_0 \neq \alpha_2$). There still exist certain values of the energy for which the transmission coefficient has a more or less sharp maximum, though smaller than 1. This phenomenon of 'selective transmission' seems to give a key to the explanation of the so-called *Ramsauer effect*, which consists in an abnormally high transparency of certain atoms (and molecules) for slow cathode rays.

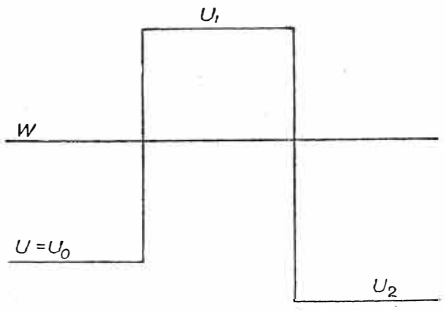

FIG. 10

2. *Transition through a Potential Energy Mountain (Tunnel Effect).* We shall now turn to the investigation of the reflection and transmission of a particle in a field of force which may be described as the *reverse* of the preceding one, with a 'mountain' or 'peak' instead of a valley. The potential energy-levels on both sides of the peak will be supposed to be different, so that the U-curve will consist of the following three parts: $U = U_0$ for $x < 0$ $(s = 0)$, $U = U_1 > U_0$ for $0 \leqslant x \leqslant a$ $(s = 1)$, and $U = U_2$ $(s = 2)$ for $x > a$ (Fig. 10). We must distinguish two cases according as the total energy W is smaller or larger than $U_{\max} = U_1$. If $W < U_1$, the penetration of an incident particle to the other side of the mountain is a result which could not occur according to classical mechanics and is therefore a characteristic consequence of wave mechanics. It may be mentioned that a quite similar phenomenon can take place and has actually been observed in the realm of optics. If two parallel glass plates are placed at a very small distance from each other—of the order of magnitude of the wave-length of the incident light—and if the angle of incidence on the inside surface of one of the plates corresponds to total reflection, total reflection will *not* take place, but a fraction of the incident energy will be transmitted through the second plate—no transmission taking place if it is removed (Fig. 11).

The intensity of the transmitted light decreases exponentially with the distance between the contiguous surfaces of the two plates. This optical phenomenon differs from the transmission of material waves—or particles—through an energy mountain only because such particles can be totally reflected at normal incidence if, as we now suppose, $W < U_1$.

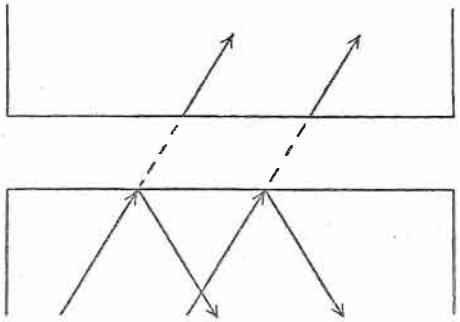

FIG. 11

Putting $\alpha_1 = i\beta_1$ and $A_2'' = 0$, i.e.

$$\psi_0 = A_0' e^{i\alpha_0 x} + A_0'' e^{-i\alpha_0 x}, \qquad \psi_1 = A_1' e^{\beta_1 x} + A_1'' e^{-\beta_1 x}, \qquad \psi_2 = A_2 e^{i\alpha_2 x},$$

we obtain from the boundary conditions (75 b)

$$A_0' + A_0'' = A_1' + A_1'', \qquad i\alpha_0(A_0' - A_0'') = \beta_1(A_1' - A_1'')$$

for $x = 0$, and

$$A_1' e^{\beta_1 a} + A_1'' e^{-\beta_1 a} = A_2 e^{i\alpha_2 a}, \qquad \beta_1(A_1' e^{\beta_1 a} - A_1'' e^{-\beta_1 a}) = i\alpha_2 A_2 e^{i\alpha_2 a}$$

for $x = a$. Solving these equations in the same way as in the preceding case we easily find

$$\begin{aligned} A_0' &= \tfrac{1}{4} A_2 e^{i\alpha_2 a}\left[\left(1+\frac{\beta_1}{i\alpha_0}\right)\left(1+\frac{i\alpha_2}{\beta_1}\right)e^{-\beta_1 a} + \left(1-\frac{\beta_1}{i\alpha_0}\right)\left(1-\frac{i\alpha_2}{\beta_1}\right)e^{\beta_1 a}\right], \\ &= \tfrac{1}{4} A_2 e^{i\alpha_2 a}\left[\left(1+\frac{\alpha_2}{\alpha_0}\right)(e^{\beta_1 a} + e^{-\beta_1 a}) + i\left(\frac{\beta_1}{\alpha_0} - \frac{\alpha_2}{\beta_1}\right)(e^{\beta_1 a} - e^{-\beta_1 a})\right], \\ &= \tfrac{1}{2} A_2 e^{i\alpha_2 a}\left[\left(1+\frac{\alpha_2}{\alpha_0}\right)\cosh\beta_1 a + i\left(\frac{\beta_1}{\alpha_0} - \frac{\alpha_2}{\beta_1}\right)\sinh\beta_1 a\right]. \end{aligned}$$

Hence $$|A_0'|^2 = \tfrac{1}{4}|A_2|^2\left[\left(1+\frac{\alpha_2}{\alpha_0}\right)^2 \cosh^2\beta_1 a + \left(\frac{\beta_1}{\alpha_0} - \frac{\alpha_2}{\beta_1}\right)^2 \sinh^2\beta_1 a\right],$$

and, according to (76 c),

$$D = \frac{4\alpha_0\alpha_2}{(\alpha_0+\alpha_2)^2\cosh^2\beta_1 a + \left(\beta_1 - \dfrac{\alpha_0\alpha_2}{\beta_1}\right)^2 \sinh^2\beta_1 a}. \tag{80}$$

When $a = 0$ this expression reduces to that obtained earlier in § 12. In the other limiting case of very large values of a, or more exactly of $\beta_1 a$, we can write approximately

$$D = \frac{16\alpha_0\alpha_2}{(\alpha_0+\alpha_2)^2+\left(\beta_1-\frac{\alpha_0\alpha_2}{\beta_1}\right)^2} e^{-2\beta a}. \tag{80 a}$$

The expression for D when $W > U_1$ can be obtained without repeating the above calculation merely by replacing β_1 in (80) by $i\alpha_1$. We thus get

$$D = \frac{16\alpha_0\alpha_2}{(\alpha_0+\alpha_2)^2\cos^2\alpha_1 a+\frac{1}{\alpha_1^2}(\alpha_1^2+\alpha_0\alpha_2)^2\sin^2\alpha_1 a}. \tag{80 b}$$

An interesting and important feature of the above expressions for D is their *symmetry with respect to* α_0 *and* α_2. This symmetry means that the transmission probability for a particle with a given energy W is *the same for the two opposite directions of motion through the field of force considered*. This result was noted before in § 12 in connexion with a simpler type of U-curve treated there. It can easily be generalized for transmission through any field of force, provided such transmission is possible ($W > U^{\pm}$), and represents a particular case of the general *reversibility law* of wave mechanics, concerning the equal probability of opposite processes, a law which will be established later.

The case $W > U_1$ does not present any particular physical interest—since it contains nothing new in principle—whereas the opposite case ($W < U_1$) throws light on some important physical phenomena which were utterly inexplicable on the basis of classical mechanics or even by Bohr's theory.

Millikan and other investigators have shown that very strong electric fields set free electrons from cold metals in the most perfect vacuum. The remarkable thing in this phenomenon is the dependence of the electron current strength J on the electric field strength E—a dependence which, with sufficient accuracy, is represented by the function $e^{-b/E}$. Fowler and Nordheim have derived this dependence theoretically on the basis of wave mechanics, by assuming for the motion of a free electron in a metal a crude model of the same sort as that used above in connexion with the hydrogen atom (Fig. 8). The model has a 'valley' extending on one side to infinity, since the thickness of the metal does not play any role in the present problem. Thus the potential-energy curve in the absence of an external electric field can be defined by $U = -V$ for $x < 0$ and $U = 0$ for $x > 0$. In the presence of a field

E acting at the surface of the metal ($x = 0$) and pulling the electron, when it reaches the surface, outwards (i.e. in the positive x-direction) the horizontal external part of the graph U is replaced by an inclined straight line $U(x) = -eEx$, whose slope is proportional to the force eE acting on the electron. The problem is thus reduced to the integration of the equations

$$\frac{d^2\psi_0}{dx^2} + \frac{8\pi^2 m}{h^2}(W+V)\psi_0 = 0 \qquad (x < 0),$$

$$\frac{d^2\psi_1}{dx^2} + \frac{8\pi^2 m}{h^2}(W+eEx)\psi_1 = 0 \qquad (x > 0),$$

with continuity conditions $\psi_0 = \psi_1$ and $d\psi_0/dx = d\psi_1/dx$ for $x = 0$, and the proper boundary conditions at $x = \pm\infty$. For the approximate determination of the function $J(E)$ we can replace the U-peak (Fig. 12)

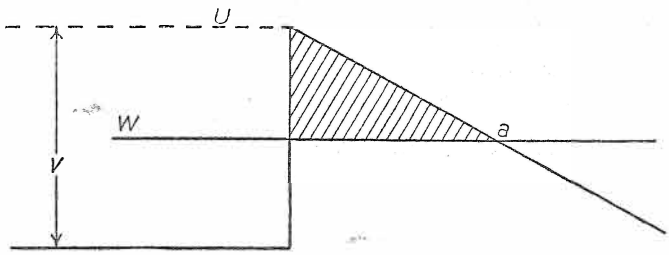

FIG. 12

corresponding to the presence of the external electric field by a rectangular region of the kind considered above (Fig. 10) with

$$\alpha_0^2 = \frac{8\pi^2 m}{h^2}(W+V), \qquad \beta_1^2 = -\frac{8\pi^2 m}{h^2}W \quad (W < 0);$$

$$\alpha_2^2 = 0 \text{ (or any const. } > 0), \qquad a = -\frac{W}{eE} = \frac{|W|}{eE}.$$

This value of a is equal to the distance from the metal surface to the place where the 'kinetic energy' of the emerging cathode rays first assumes a positive value. Since the current must be proportional to the 'permeability coefficient' of the confined region, we obtain by (80 a)

$$J \sim e^{-b/E}, \tag{81}$$

with

$$b = \frac{2\pi}{eh}\sqrt{(2m)}|W|^{\frac{3}{2}}. \tag{81 a}$$

A second phenomenon very similar in principle to the one just considered is the decay of radioactive atoms with emission of α-rays. Here the α-particles contained in the positive nucleus play the part of the

electrons in the metal, while the Coulomb field which is produced by the nucleus itself in the surrounding space replaces the electric field applied externally to the metal surface. Inside the nucleus the α-particle must of course be held by some attractive force (whose nature and origin is unknown as yet) just as the free electrons responsible for the conductivity of the metal are retained inside the metal. The permeability coefficient of the nucleus for an α-particle determines in this case the probability of its emission, that is, the radioactive decay constant of the element under consideration. By a simple calculation Gamow, and independently Condon and Gurney, succeeded

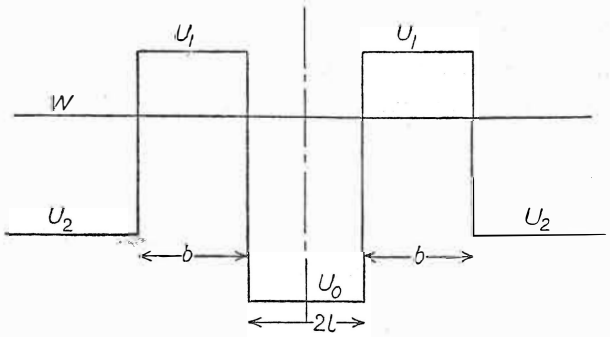

FIG. 13

in determining the dependence of this constant on the energy of the emitted α-particle, in agreement with the empirical formula of Geiger and Nuttal.

We shall use for this calculation a simplified uni-dimensional model of the nucleus, formed by a field of force which is obtained by reflecting Fig. 10 with respect to the ordinate axis. We get in this way a symmetrical curve of the type shown in Fig. 13 but of a simplified rectangular shape, specified by the following data: $U = U_0$ for $-l < x < l$, $U = U_1 > U_0$ for $l < x < l+b$ $(s = 1)$, and $-(l+b) < x < -l$ $(s = -1)$ and $U = U_2 < U_1$ for $x > l+b$ $(s = 2)$ and $x < -(l+b)$ $(s = -2)$.

We shall assume the motion of the α-particle inside the nucleus, i.e. within the valley U_0, to be approximately the same as if the width (b) of the surrounding potential mountains were infinite, in which case the particle would be unable to escape from the nucleus. Let us imagine instead of a single particle a number of copies, measured by the integral $\int_{-l}^{+l} |\psi_0|^2 \, dx = N$, say. The function ψ_0 can be considered as the superposition of two waves travelling in opposite directions and reflected at

the walls $x = \pm l$. Actually the reflection is not complete so long as U_1 and b are finite. The number of particles escaping through one of the walls in unit time is equal to the product of the current density I of the stream directed at this wall† by the permeability coefficient D. Now I is equal to the product of the number of particles per unit-length in the *incident* stream, $\frac{1}{2}\frac{N}{2l} = \frac{1}{2}\frac{N}{a}$ where $a = 2l$ is the width of the valley, by the velocity $v_0 = \frac{\alpha_0}{2\pi m} = \sqrt{\left\{\frac{2}{m}(W - U_0)\right\}}$. The factor $\frac{1}{2}$ arises from the fact that one-half of the particles are moving in one direction and the other half in the opposite direction. Adding the currents through the two walls, we get for the rate of change of N with time the equation

$$\frac{dN}{dt} = -\frac{N}{a} v_0 D; \qquad \text{or} \qquad N = \text{const. } e^{-\gamma t},$$

i.e. the equation of radioactive decay with the following expression for the decay constant

$$\gamma = \frac{v_0 D}{a}, \tag{82}$$

or approximately

$$\gamma \cong \frac{\alpha_0}{2\pi m a} e^{-2\beta_1 b}. \tag{82 a}$$

3. *Quasi-stationary States and Leaking Effect.* The preceding calculation was based upon the assumption that the motion of the α-particle within the nucleus was unaffected by the permeability of the walls limiting it, i.e. that this motion remains quantized in the same way as in the case of the hydrogen-atom model discussed above (with $U_0 = -V$, $U_1 = 0$, and $b = \infty$). We shall now consider the influence of this permeability on the motion.

It is clear from the general discussion in the preceding section that there will be, strictly speaking, no quantization at all so long as the energy W is larger than U_2. This means that under the condition $W > U_2$ all values of the energy are allowed, just as in the case $W > U_1$. The stationary states of the particle can be represented accordingly by standing waves inside the valley and in the external regions; in the intermediate regions (corresponding to the 'mountains') with $W < U_1$ we shall, however, have vibrations non-periodic with respect to x, characteristic of total reflection.

We can limit ourselves as before to the case of standing waves sym-

† This density is referred to the whole area of the wall, which can be taken as unity, so that the volume of the internal region of the nucleus is equal to a.

metrical and antisymmetrical with respect to the origin. For symmetrical waves we have

$$\psi_0 = A_0 \cos \alpha_0 x, \qquad \psi_1 = A_1' e^{\beta_1 x} + A_1'' e^{-\beta_1 x}, \qquad \psi_2 = A_2 \cos(\alpha_2 x + \phi_2).$$

ψ_{-1} and ψ_{-2} need not be considered separately since they are obtained from ψ_1 and ψ_2 by simply changing the sign of x. The continuity conditions give the relations:

$$A_0 \cos \alpha_0 l = A_1' e^{\beta_1 l} + A_1'' e^{-\beta_1 l}, \qquad -\alpha_0 A_0 \sin \alpha_0 l = \beta_1 (A_1' e^{\beta_1 l} - A_1'' e^{-\beta_1 l}),$$
$$A_1' e^{\beta_1 (l+b)} + A_1'' e^{-\beta_1 (l+b)} = A_2 \cos[\alpha_2 (l+b) + \phi_2],$$
$$\beta_1 (A_1' e^{\beta_1 (l+b)} - A_1'' e^{-\beta_1 (l+b)}) = -\alpha_2 A_2 \sin[\alpha_2 (l+b) + \phi_2].$$

The interesting point in our problem is the ratio $\frac{A_2}{A_0}$ or $\left|\frac{A_2}{A_0}\right|^2$. It can be obtained by solving the first two equations with respect to A_1' and A_2', substituting the results in the second two, and finally squaring and adding them after dividing the last one by α_2. This gives

$$4\frac{A_2^2}{A_0^2} = \left[\left(\cos \alpha_0 l - \frac{\alpha_0}{\beta_1} \sin \alpha_0 l\right) e^{\beta_1 b} + \left(\cos \alpha_0 l + \frac{\alpha_0}{\beta_1} \sin \alpha_0 l\right) e^{-\beta_1 b}\right]^2 + \left. + \frac{\beta_1^2}{\alpha_2^2}\left[\left(\cos \alpha_0 l - \frac{\alpha_0}{\beta_1} \sin \alpha_0 l\right) e^{\beta_1 b} - \left(\cos \alpha_0 l + \frac{\alpha_0}{\beta_1} \sin \alpha_0 l\right) e^{-\beta_1 b}\right]^2 \right\}. \quad (83)$$

The minimum of this expression corresponds approximately to

$$\cos \alpha_0 l - \frac{\alpha_0}{\beta_1} \sin \alpha_0 l = 0; \qquad \text{i.e.} \qquad \tan \alpha_0 l = \frac{\beta_1}{\alpha_0},$$

and the maximum to

$$\cos \alpha_0 l + \frac{\alpha_0}{\beta_1} \sin \alpha_0 l = 0; \qquad \text{i.e.} \qquad \tan \alpha_0 l = -\frac{\beta_1}{\alpha_0}.$$

The degree of approximation, i.e. the accuracy with which the respective values of the energy W are determined by these equations, is the higher, the larger the value of b, or rather, the product $\beta_1 b$. Assuming $\beta_1 b$ to be large enough, we get

$$(A_2/A_1)^2_{\min} = \left(1 + \frac{\beta_1^2}{\alpha_2^2}\right) \cos^2 \alpha_0 l \; e^{-2\beta_1 b}$$

or, since $\cos^2 \alpha_0 l = 1/(1 + \tan^2 \alpha_0 l)$ and $\tan \alpha_0 l = \beta_1/\alpha_0$,

$$\left(\frac{A_2}{A_1}\right)^2_{\min} = \frac{1 + \beta_1^2/\alpha_2^2}{1 + \beta_1^2/\alpha_0^2} e^{-2\beta b}. \quad (83\,a)$$

In a similar way we get

$$\left(\frac{A_2}{A_1}\right)^2_{\max} = \frac{1 + \beta_1^2/\alpha_2^2}{1 + \beta_1^2/\alpha_0^2} e^{+2\beta b}. \quad (83\,b)$$

A comparison with equation (78 a) shows that the states—and in particular the energies—corresponding to the minimum value of A_2/A_1 coincide with the (symmetrical) quantized states which would be obtained if the width b of the two mountains or 'walls' were infinite, their height U_1 remaining the same. With a finite value of b there is a very small 'leaking' through the walls into the external regions, the rate of which is characterized by the ratio A_2/A_1. This leaking does not actually appear because the outgoing waves with which it should be associated are superposed in the outward regions with incoming waves representing compensating streams. This superposition forms standing waves $\psi_{\pm 2}$. It is, however, easy to calculate what would be the rate of leak (or 'decay') if these compensating streams were not present, by assuming that $\psi_{\pm 2}$ reduces to the outgoing waves only, with an amplitude $A'_{\pm 2} = A_2/\sqrt{2}$. This assumption is not exactly correct, because the outgoing waves are due not only to transmission from the inside, but also to a partial reflection of the waves incident from the outside (i.e. the incoming waves) which we are now supposing to be removed. Leaving this out of account, we get for the outward stream per unit time through both walls $|A_2|^2 v_2$. Dividing this by the product $A_0^2 a$, which represents the relative number of copies of the particle remaining in the internal region (or in other words the probability of finding the particle considered within this region), we get for the leak or decay constant the expression

$$\gamma = \left(\frac{A_2}{A_0}\right)^2_{\min} \frac{v_2}{a} = \frac{\alpha_2}{2\pi m a} \frac{1+\beta_1^2/\alpha_2^2}{1+\beta_1^2/\alpha_0^2} e^{-2\beta b}. \tag{83 c}$$

This differs slightly from the expression (82) or (82 a) found before, because in deriving (83 c) we have identified the outgoing components of the external standing wave $\psi_{\pm 2}$ with the transmitted waves, and have made no allowance for the reflection of the incoming waves. It is interesting to note that the minima of A_2/A_0 are the smaller and the maxima the larger, the larger the width of the transition zone b (or its height U_1). For states corresponding to the maximum values of A_2/A_0, i.e. the minimum values of A_0/A_2, the preceding argument can be reversed and the leaking from the inside to the outside replaced by a leaking of practically the same amount in the opposite direction. We can express these results† by saying that when the energy of the particle is approximately equal to that of a state which would have been stationary if

† And also the entirely similar results which are obtained in the case of motion of the type antisymmetrical with respect to the origin.

$b = \infty$, the particle displays a strong preference to remain 'inside', i.e. in the region of minimum potential energy, in a state very similar to that just mentioned, differing from it by a certain tendency to escape finally to the outside. When the particle has an energy defined by $\tan \alpha_0 l = -\beta_1/\alpha_0$, which lies approximately in the middle of the interval between two consecutive levels, corresponding to the above stationary states, the particle displays on the contrary a tremendous repugnance for entering the internal region, preferring to remain outside. This repugnance is not limited to the states corresponding to $(A_2/A_0)^2$ = maximum; it becomes manifest when the ratio A_2/A_0 is larger than 1. Thus the balance is in favour of the interior for very narrow intervals of the energy only. These intervals, which are practically equal to zero if $\beta_1 b$ is large, correspond to a variation of the ratio A_2/A_0 from its minimum value to 1. The width of these intervals can be defined as the width of the 'quasi-quantized' energy-levels corresponding to 'quasi-stationary' states of the particle, and can easily be estimated by means of the equation (83). Putting $A_2/A_0 = 1$, and dropping the terms in $e^{-\beta_1 b}$ which contribute practically nothing to the result, we get

$$4 = \left(1+\frac{\beta_1^2}{\alpha_0^2}\right) e^{2\beta_1 b}\left(\cos\alpha_0 l - \frac{\alpha_0}{\beta_1}\sin\alpha_0 l\right)^2,$$

or

$$\left|\cos\alpha_0 l - \frac{\alpha_0}{\beta_1}\sin\alpha_0 l\right| = \frac{2}{\sqrt{(1+\beta_1^2/\alpha_0^2)}}\, e^{-\beta_1 b}.$$

Since this is a very small quantity we can replace it by the product $\blacktriangle W \frac{\partial}{\partial W}\left(\cos\alpha_0 l - \frac{\alpha_0}{\beta_1}\sin\alpha_0 l\right)$, the derivative with respect to W being taken for that value of $W = W_n$ which corresponds to

$$\cos\alpha_0 l - \frac{\alpha_0}{\beta_1}\sin\alpha_0 l = 0.$$

With the help of the relations $\frac{d\alpha_0}{dW} = \frac{4\pi^2 m}{h^2\alpha_0}$ and $\frac{d\beta_1}{dW} = \frac{4\pi^2 m}{h^2\beta_1}$, we find, after a few simplifications,

$$\Delta W \frac{4\pi^2 m}{h^2\sqrt{(1+\beta_1^2/\alpha_0^2)}}\left[\frac{l}{\beta_1}\left(1+\frac{\beta_1^2}{\alpha_0^2}\right)+\frac{1}{\alpha_0^2}-\frac{1}{\beta_1^2}\right] = \frac{2}{\sqrt{(1+\beta_1^2/\alpha_0^2)}}\, e^{-\beta_1 b}.$$

If β_1 is large compared with α, this reduces approximately to

$$\Delta W \cong \frac{h^2\alpha_0^2}{4\pi^2 m l \beta_1}\, e^{-\beta_1 b}. \tag{83 d}$$

It should be noticed that the quantity $h^2\alpha_0^2/(8\pi^2 m)$ is equal to the kinetic

energy of the particle in the internal region for the quasi-stationary state in question.

The width of the energy-levels, corresponding to the quasi-stationary states, can be defined in what seems to be a more natural way as that value of ΔW for which the ratio A_2^2/A_1^2 increases *twice*, say, compared to its minimum value (by analogy with the usual definition of the width of a spectrum line). We thus get instead of (83 d) a width given approximately by the condition $\Delta(\cos\alpha_0 l - \frac{\alpha_0}{\beta_1}\sin\alpha_0 l) = e^{-2\beta_1 b}$, that is,

$$\Delta W = \frac{h^2\alpha_0^2}{4\pi^2 m l \beta_1} e^{-2\beta_1 b}, \tag{83 e}$$

which is smaller than the value given in (83 d) by the factor $e^{-\beta_1 b}$.

The results of the preceding analysis can be summed up as follows. If in addition to attractive forces, tending to hold a particle within a limited region, there are repulsive forces outside of this region, allowing the particle to escape with positive kinetic energy through a transition zone where its kinetic energy must be negative, then the particle will be able to move inside nearly in the same way and with nearly the same energy as if these repulsive forces were absent. Their influence will be manifested in a 'broadening' of the (quasi) quantized energy-levels and in a probability of escaping outside, proportional to this broadening [or to its square according to (83 d)].

A similar state of affairs is met with in many physical phenomena, of which radioactive disintegration is one of the simplest examples. The extraction of electrons from a metal by an electric field applied to its surface is another. In an elementary form the latter effect is found in the ionization of an atom produced by an electrostatic field (see Part III). A similar condition is met with in the case of many molecules which are dissociated 'spontaneously' (i.e. without any external perturbation) after having been first brought to a state called by V. Henri a state of *predissociation*. We can use this term to indicate all the quasi-stationary or quasi-quantized states that have been studied above.

4. *Motion of an Electron in a Crystal-Lattice.* In conclusion we shall apply the above method to a somewhat more complicated problem, namely, the motion of an electron in what can be considered as the crudest linear model of a crystalline body. This model is specified by a potential energy function $U(x)$ assuming alternately the constant values U_a over a range a and $U_b > U_a$ over the next range b, $a+b = c$ being the 'lattice constant'. The function $U(x)$ satisfies the periodicity

condition $U(x+c) = U(x)$. The crystal or rather the toothed wall representing it (Fig. 14) will be considered as infinite.

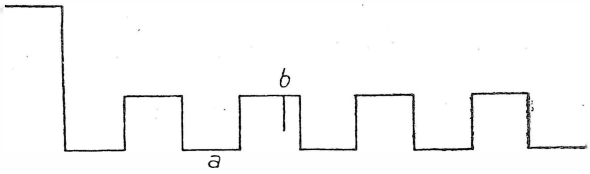

FIG. 14

For each period of the lattice extending from $x = nc$ to $x = (n+1)c$, where n is a whole number, the wave function can be defined as the sum

$$\left.\begin{aligned} \psi_n^a &= A'_n e^{i\alpha x} + A''_n e^{-i\alpha x} \qquad (cn \leqslant x \leqslant cn+a) \\ \psi_n^b &= B'_n e^{\beta x} + B''_n e^{-\beta x} \qquad (cn+a \leqslant x \leqslant c(n+1)) \end{aligned}\right\}, \tag{84}$$

where α and β are expressed in the usual way†; β is assumed to be real, that is, larger than the energy W. If this is not the case we have only to replace β by $i\beta$. The functions ψ_n^a and ψ_n^b must satisfy the usual boundary conditions, which give a series of equations for the determination of the relations between the coefficients A and B for the successive regions.

These equations, since their number is infinite, are rather difficult to handle. A simple method for their solution has been devised by E. Hill. A still simpler procedure has been indicated by R. Kronig and W. Penney. It is based on the fact that particular solutions of the type (84) can be written in the following form, valid for all values of x,

$$\psi = C(x)e^{i\gamma x}, \tag{84a}$$

where C is a function periodic with regard to the lattice, i.e. satisfying the condition

$$C(x+c) = C(x),$$

while γ is a certain suitably chosen parameter. The motion associated with the function (84a) can be pictured (if γ is real and positive) as a uniform translation in the positive direction with an average velocity $\bar{v} = \hbar\gamma/(2\pi m)$, combined with a sort of vibration which is described by the amplitude function $C(x)$.

It should be remarked that the general solution of our problem for an arbitrary value of the ratio A'/A'' in (84) can be represented by the superposition of two functions of the type (84a) with opposite values of the parameter γ.

† $\alpha = \frac{2\pi}{h}\sqrt{\{2m(W-U_a)\}}$, $\beta = \frac{2\pi}{h}\sqrt{\{2m(U_b-W)\}}$.

The function $C(x)$ is given by

$$C(x) = A'_n e^{i(\alpha-\gamma)x} + A''_n e^{-i(\alpha+\gamma)x}$$

for $cn \leqslant x \leqslant cn+a$, and

$$C(x) = B'_n e^{(\beta-i\gamma)x} + B''_n e^{-(\beta+i\gamma)x}$$

for $cn+a \leqslant x \leqslant c(n+1)$.

Writing down the periodicity condition for the function $C(x)$ and its derivative for the points $x = cn$ and $x = c(n+1)$, we get:

$$A'_n e^{i(\alpha-\gamma)cn} + A''_n e^{-i(\alpha+\gamma)cn} = B'_n e^{(\beta-i\gamma)c(n+1)} + B''_n e^{-(\beta+i\gamma)c(n+1)}$$

$$i(\alpha-\gamma)A'_n e^{i(\alpha-\gamma)cn} - i(\alpha+\gamma)A''_n e^{-i(\alpha+\gamma)cn} = (\beta-i\gamma)B'_n e^{(\beta-i\gamma)c(n+1)} - (\beta+i\gamma)B''_n e^{-(\beta+i\gamma)c(n+1)}.$$

To these equations we must add the two equations

$$A'_n e^{i\alpha(cn+a)} + A''_n e^{-i\alpha(cn+a)} = B'_n e^{\beta(cn+a)} + B'_n e^{-\beta(cn+a)}$$

$$i\alpha(A'_n e^{i\alpha(cn+a)} - A''_n e^{-i\alpha(cn+a)}) = \beta(B'_n e^{\beta(cn+a)} - B''_n e^{-\beta(cn+a)}$$

resulting from the continuity conditions for the function ψ at $x = cn+a$. Eliminating the quantities A'_n, A''_n, B'_n, B''_n between these four equations, we obtain the following relation:

$$\cos \gamma c = \frac{\beta^2-\alpha^2}{2\alpha\beta} \sin \alpha a \sinh \beta b + \cos \alpha a \cosh \beta b. \qquad (84\,b)$$

This relation leads to real values of γ if the right-hand side is numerically smaller than 1. In the contrary case the parameter γ turns out to be imaginary, which means that the electron cannot travel through the lattice in the way indicated above, the waves representing its motion having the same character as in the case of total reflection. As a matter of fact, imaginary values of γ correspond to *selective reflection* of the electron waves, a selective reflection which becomes manifest in experiments on the diffraction of cathode rays by crystal lattices.

In the case of Röntgen rays falling normally on a crystal lattice with a constant spacing c between the planes, parallel to the surface, the condition for selective reflection is given by Bragg's formula

$$2c = n\lambda,$$

where λ is the wave-length and n an integer, giving the order of reflection. The equation (84 b) can be considered to be the exact form of Bragg's formula for cathode rays travelling in our simplified unidimensional model of a crystal lattice, the 'effective' wave-length being connected with the parameter γ in the same way as for free electrons

$$\lambda_{\text{eff}} = 2\pi/\gamma.$$

In fact, putting $\cos \gamma c = \pm 1$, which is the limiting form of the condi-

tion for selective reflection, we get $\gamma c = n\pi$, that is, $\lambda_{\text{eff}} = 2c/n$, which is just Bragg's formula. It should be remembered, however, that the notion of effective wave-length loses all meaning when γ becomes imaginary.

These imaginary values of γ are obtained for values of the energy W lying in certain regions of finite width, which, as can easily be shown, is the smaller the larger the energy. In the limiting case of very large values of the energy when α reduces to $\frac{2\pi}{h}\sqrt{(2mW)}$ and β to $\frac{2\pi}{h}\sqrt{(-2mW)}$, these 'forbidden' regions contract to points defined by

$$W = \frac{n^2h^2}{8mc^2}, \tag{84 c}$$

which is the kinetic energy of a practically free electron, moving with a velocity corresponding to the wave-length $\lambda = 2c/n$. In the opposite extreme case of very small energies (β very large) the *allowed* regions (i.e. regions of the energy spectrum, corresponding to real values of γ) reduce to points defined by

$$W = \frac{n^2h^2}{8ma^2},$$

which is the kinetic energy of an electron confined to move between two impenetrable potential barriers at a distance a apart. The electron thus becomes practically *bound* in one definite hollow in the lattice.

The dependence of $\cos\gamma c$ as defined by (84 b) upon the energy W with $\alpha = \frac{2\pi}{h}\sqrt{\{2m(W-U_a)\}}$ and $\beta = \frac{2\pi}{h}\sqrt{\{2m(U_b-W)\}}$ is illustrated by Fig. 15, taken from the paper by Kronig and Penney,† to which the

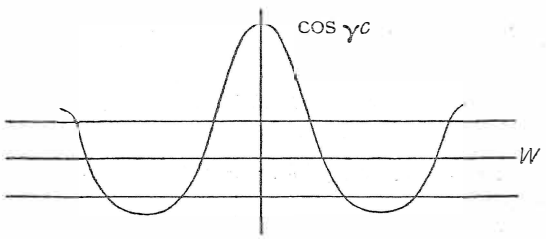

FIG. 15

reader is referred for further particulars. The allowed values of the energy correspond to the portions of the curve between the straight lines $\cos\gamma c = \pm 1$. The preceding considerations refer to the case of an *unlimited* lattice. In order to give a consistent theory of the reflection

† *Proc. Royal Soc.* **130**, p. 502, 1931. The authors use a somewhat simplified function, corresponding to $b \to 0$, and $\beta \to \infty$ with $\beta^2 b$ = finite. Cf. also Hill's paper, *Phys. Rev.* **37**, 785, 1931.

and transmission of cathode rays through a crystal, we should consider a lattice bounded at least on one side (see preceding figure) by a sudden jump of the potential energy from the value U_a, say to U_∞. Only those electrons will then be able to escape from the lattice for which the energy W is larger than U_∞. We shall not discuss this case here, but merely mention the following interesting feature of such a limited lattice, which has been pointed out by Ig. Tamm. It is in fact possible to imagine that an electron is bound *at the surface* of the body, being unable to escape outside because of insufficient kinetic energy, i.e. because of the *total reflection* of the waves representing it in the external region, and at the same time being unable to move inside, even if possessing sufficient kinetic energy, because of the *selective reflection* of the associated waves in the internal region, connected with an imaginary value of γ. The energy-levels corresponding to such 'surface-bound' electrons (which probably account for certain phenomena observed in insulating crystals) form a discrete spectrum and can be calculated from the continuity condition for the function ψ represented by $Be^{+\beta x}$ for $x < 0$, with $\beta = \frac{2\pi}{h}\sqrt{\{2m(U_\infty - W)\}}$, and by $C(x)e^{-|\gamma|x}$ for $x > 0$ with γ purely imaginary.

17. Superposition of Different Quantized States: Wave Mechanical Theory of Light Emission

Classical mechanics and Bohr's atomic theory do not recognize such a thing as the coexistence of two different (alternative) states of motion. It is, e.g., impossible to admit in those theories that a hydrogen atom could be simultaneously in two or more different quantized states. In wave mechanics, however, we meet with a situation which looks like a recognition of the possibility of such coexistence, expressed in the form of the *superposition* (or interference) *principle*. The validity of this principle is clearly demonstrated by the phenomena of interference (or diffraction) of cathode rays, and we have used it in the foregoing sections in superposing travelling waves of the same frequency to form standing waves or in superposing waves of neighbouring frequencies to form wave groups. It seems natural to make a further generalization of the superposition principle by extending it to wave functions of *different* frequencies, representing stationary states of motion with quite different values of the energy. If we forget for a moment the corpuscular properties of matter and think of the material waves as of a kind of vibrations similar to elastic or optical vibrations, the superposition of

vibrations of different frequency looks certainly just as natural as the superposition of vibrations of the same frequency.

Now it is easy to see that in wave mechanics such a superposition *does not imply the coexistence of the different states of motion associated with the corresponding harmonic vibrations.* In fact, from the point of view of the probability interpretation of the wave-corpuscular parallelism, the addition of several wave functions ψ_1, ψ_2,... must correspond to the addition of probabilities in classical probability theory, the probabilities being replaced by their amplitudes. Now in the classical probability theory this operation of addition only has a meaning when these probabilities refer to *alternative* events, i.e. events which are mutually exclusive and therefore cannot coexist. The same must be true in the new probability theory, i.e. wave mechanics. In fact it has no meaning to add a wave function to itself, for the result, or more generally the product of the original wave function by any constant factor, represents exactly the same thing, i.e. the same state of motion, as the original wave function. The sum

$$\psi = \psi_1 + \psi_2 + \psi_3 + \ldots$$

thus has a definite meaning only when the wave functions (probability amplitudes) refer to different or rather alternative states of motion. So far as the meaning of ψ is concerned, it is immaterial whether these alternative states have the same or different energies, that is, whether the harmonic vibrations represented by the separate wave functions have the same or different frequencies.

Although the multiplication of each of the functions ψ_n ($n = 1, 2, \ldots$) by an arbitrary constant c_n does not alter its meaning, the sum

$$\psi = \sum_n c_n \psi_n \tag{85}$$

has a meaning which depends upon the *relative* values of the different coefficients c_n, as is clearly seen if we take the extreme case, when some of these coefficients are zero. It must be emphasized that it is only these relative values, i.e. the ratios between the coefficients c_n, which are significant and not the absolute values; a multiplication of the whole sum by an arbitrary factor does not alter the meaning of the function ψ.

This ambiguity in the absolute value of ψ, or of the elementary wave functions

$$\psi_n(x, y, z, t) = \psi_n^0(x, y, z) e^{-i 2\pi \nu_n t}, \tag{85 a}$$

is usually removed by means of a so-called '*normalizing condition*'. This condition is similar—and in fact very nearly related—to the condition,

used for the same purpose in classical probability theory, which consists in setting the maximum value of probability, amounting to certainty, equal to 1. This normalizing condition for probabilities is usually introduced in connexion with alternative events which are taken into consideration as possible (or probable) issues; the relative values of the probabilities of these events being known, their absolute values are fixed by 'normalizing' their sum to 1.

The 'normalization' of a function ψ_n is achieved on the same basis. The product $|\psi_n|^2\,dV$ ($dV = dxdydz$) being the relative probability for finding the particle in a certain stationary state of motion ψ_n, in the volume-element dV, the probability of finding it within some finite region of space will be measured by the integral $\int |\psi_n|^2\,dV$ taken over the volume V of this region. This result implies that the location of the particle in different portions of V (while it is in the same state of motion) are alternative events, to which the classical additivity law of probabilities applies. An interference of probabilities, requiring their replacement by probability amplitudes, can take place only in connexion with the same element dV.

If the integral $\int |\psi_n|^2\,dV = \int |\psi_n^0|^2\,dV$ remains finite when V tends to infinity, i.e. when the integration is extended over the whole space, it can be normalized to 1, that is, we can put

$$\int \psi\psi_n^*\,dV = 1 \tag{85 b}$$

to express the fact that the particle must *certainly* be somewhere. The function ψ_n or rather ψ_n^0 is in this case said to be 'quadratically integrable'. Such quadratically integrable wave functions represent stationary motions of the particle in a 'bound' condition, i.e. with an energy W smaller than the potential energy at infinite distance. The fact that this condition implies an exponential decrease of the wave functions with increase of distance, a decrease which ensures the convergence of the integral $\int \psi\psi^*\,dV$, has been proved in § 14 for unidimensional motion; the result can easily be extended to the general case. It was shown at the same time that the 'bound condition' implies the existence of 'discrete' or 'quantized' stationary states, corresponding to discrete values of the energy W_n. We thus see that the normalization condition (85 b) can be applied to quantized states of a bound particle only. In the case of wave functions representing the stationary states of a 'free' particle, that is, a particle capable of escaping to infinity, the integral $\int \psi_n\psi_n^*\,dV$ diverges, so that the normalization condition (85 b) is inapplicable. It can be conveniently replaced in this

case, as will be shown later on (Part II), by the condition

$$\iint \psi_n^0(x,y,z)\psi_n^{0*}(x',y',z')\,dV dV' = 1 \quad (dV' = dx'dy'dz'). \tag{85 c}$$

We shall confine ourselves in what follows to quantized motions and use accordingly the normalization condition (85 b). Under this condition the coefficients c_n in the equation (85) can be considered, from the wave point of view, as the *amplitudes* of the corresponding simple harmonic vibrations in the compounded vibration produced by their superposition. From the corpuscular point of view in conjunction with the probability interpretation they must obviously determine the relative *probabilities of finding the electron in the associated quantized* states, irrespective of its position in space.

In order to establish this connexion in a more precise way (and incidentally to establish a normalization condition for the function ψ) we must first define the probability density ρ, that is, the probability ρdV of finding the electron in the volume-element dV, for a motion associated with the compounded vibrations represented by the wave function (85). The usual definition of ρ as the square of the (real) amplitude of ψ is inapplicable here, since we have to do with a wave process which is not even approximately harmonic (as, for instance, in the case of wave groups). It seems natural, however, to generalize the original definition by redefining ρ as the square of the *modulus* (absolute value) $|\psi|$, since the modulus coincides with the amplitude in the special case of harmonic (or nearly harmonic) vibrations represented by a complex time factor $e^{-i2\pi\nu_n t}$. This generalization is in accord with the fact that ψ is a complex quantity (for the multiplication law to hold, the elementary functions ψ_n must be complex, see below), whereas ρ must obviously be a real quantity, quadratic in ψ. There are two other consequences of the assumption

$$\rho = |\psi|^2 = \psi\psi^* \tag{86}$$

(the asterisk denotes the conjugate complex quantity) which serve to justify it, and which we shall presently examine.

According to (86),

$$\psi\psi^* = \sum_m \sum_n c_m^* c_n \psi_m^* \psi_n. \tag{86 a}$$

Integrating this over the whole space, we get

$$\iint \psi\psi^*\,dV = \sum_m \sum_n c_m^* c_n \int \psi_m^* \psi_n\,dV. \tag{86 b}$$

Now it can easily be shown that

$$\iint \psi_m^* \psi_n\,dV = 0, \tag{87}$$

unless the two functions ψ_m and ψ_n are identical, in which case, according to the normalization condition (85 b), the zero should be replaced by unity. We will first prove this result for the simple model of an electron enclosed in a rectangular box, which was examined in § 11.

We have in fact, according to (37) (restoring the time factor there dropped) and (36 a),

$$\psi_n = A_n \sin\frac{\pi n_x x}{a} \sin\frac{\pi n_y y}{b} \sin\frac{\pi n_z z}{c} e^{-i2\pi\nu_n t},$$

where the index n stands for brevity for the three indices n_x, n_y, n_z. Multiplying this by

$$\psi_m^* = A_m^* \sin\frac{\pi m_x x}{a} \sin\frac{\pi m_y y}{b} \sin\frac{\pi m_z z}{c} e^{+i2\pi\nu_m t},$$

and integrating over the volume of the box $V = abc$, we obtain the product of $A_m^* A_n e^{i2\pi(\nu_m - \nu_n)t}$ by three integrals of the type

$$\int_0^\infty \sin\frac{\pi m_x x}{a} \sin\frac{\pi n_x x}{a}\, dx,$$

which is equal to zero when $m_x \neq n_x$, and to $\frac{1}{2}a$ when $m_x = n_x$.† Thus unless all the three numbers m are equal to the corresponding numbers n, the integral $\int \psi_m^* \psi_n\, dV$ vanishes. It may be remarked that in the contrary case ($m = n$) we have $\int \psi_m^* \psi_n\, dV = \frac{1}{8}abc\, A_n A_n^*$, so that the normalization condition (85 b) gives

$$|A_n|^2 = \frac{8}{V} \qquad (V = abc).$$

The general proof of (87) can be given with the help of Schrödinger's equation $\nabla^2\psi_n + \frac{8\pi^2 m_0}{h^2}(W_n - U)\psi_n = 0$ by which the functions ψ_n and the corresponding energy values are defined for an arbitrary potential energy function $U(x, y, z)$ (consistent with bounded states of motion). Multiplying this equation by ψ_m^* and subtracting the result from the product

$$\psi_n\left[\nabla^2\psi_m^* + \frac{8\pi^2 m_0}{h^2}(W_m - U)\psi_m^*\right] = 0,$$

† In fact one has

$$\int_0^a \sin\alpha x \sin\beta x\, dx = \tfrac{1}{2}\int_0^a [\cos(\alpha-\beta)x - \cos(\alpha+\beta)x]\, dx = \tfrac{1}{2}\left[\frac{\sin(\alpha-\beta)a}{\alpha-\beta} - \frac{\sin(\alpha+\beta)a}{\alpha+\beta}\right].$$

Since $\alpha = \frac{\pi m_x}{a}$ and $\beta = \frac{\pi n_x}{a}$, this is equal to zero when $m_x \neq n_x$. If $m_x = n_x$, we have, on the other hand, $\sin(\alpha-\beta)/\alpha-\beta = a$.

we get $$\psi_n \nabla^2 \psi_m^* - \psi_m^* \nabla^2 \psi_n = \frac{8\pi^2 m_0}{h^2}(W_n - W_m)\psi_m^* \psi_n.$$

Now the left-hand side of this equation can be written in the form

$$\frac{\partial}{\partial x}\left(\psi_n \frac{\partial \psi_m^*}{\partial x} - \psi_m^* \frac{\partial \psi_n}{\partial x}\right) + \frac{\partial}{\partial y}\left(\psi_n \frac{\partial \psi_m^*}{\partial y} - \psi_m^* \frac{\partial \psi_n}{\partial y}\right) + \frac{\partial}{\partial z}\left(\psi_n \frac{\partial \psi_m^*}{\partial z} - \psi_m^* \frac{\partial \psi_n}{\partial z}\right),$$

that is, in the form of the divergence of a vector whose components are represented by the terms in the brackets. If we multiply this divergence by dV and integrate over a finite volume V, we obtain a result which, according to Green's theorem, can be expressed as an integral taken over the surface limiting V with the normal component of the above vector as integrand. In the limit when the volume integration is extended over all space this surface integral vanishes—in view of the rapid decrease of both ψ_m and ψ_n, so that we obtain the equation

$$(W_n - W_m) \int \psi_m^* \psi_n \, dV = 0.$$

This reduces to $\int \psi_m^* \psi_n \, dV = 0$ if $W_n \neq W_m$. We thus see that the relation (87) is certainly fulfilled if the states m and n belong to two different values of the energy. As has been shown in connexion with the problem of the isotropic oscillator and the hydrogen atom, it is possible to have different states, i.e. states associated with different wave functions, belonging to the same energy value. For these states the relation (87) is in general not fulfilled. The existence of such 'degenerate' states is, however, always connected with a certain ambiguity in their definition, an ambiguity which has already been illustrated in the cases referred to, and which can be expressed in a general form as follows: *if two or more functions ψ_n represent stationary states with the same energy, then any linear combination of these functions also represents a stationary state with the same energy.* This result is a direct consequence of the linear character of Schrödinger's equation. It can be used to define the stationary states of a degenerate group in such a way that the condition (87) is fulfilled for them too. So far as the function $\psi = \sum c_n \psi_n$ is concerned, this is not necessary, for the sum of terms belonging to the same energy can be replaced by a single term of the same type. Turning back to the integrated equation (86 b) we see that it reduces to a sum of squares

$$\int \psi\psi^* \, dV = \sum_n |c_n|^2. \tag{88}$$

Now the integral $\int \psi\psi^* \, dV$ must obviously be interpreted as a measure of the total probability of finding the particle in any one of

the alternative states implied in the function ψ. On the other hand it is represented, according to (88), as a sum of squares of the moduli of the amplitude coefficients c_n for all the alternative states referred to. This result is in agreement with the addition theorem of classical probability theory if the quantities

$$|c_n|^2 = c_n c_n^*$$

are interpreted as the relative probabilities of finding the particle in the corresponding alternative states, *irrespective of its position in space.* They can be conveniently normalized by the condition

$$\sum_n |c_n|^2 = 1. \tag{88 a}$$

In view of the fact that the relation $\int \psi_m^* \psi_n \, dV = 0$ reduces the integral $\int \psi\psi^* \, dV$, i.e. the double sum (86 b), to a sum of squares, it is called the *orthogonality relation.* It can be regarded as the analytical expression of the alternative character of the corresponding states.

The above interpretation of the coefficients c_n was introduced by Born on the basis of the definition (86). There is, however, another still more impressive justification of this definition which has been pointed out (with a theoretically different though practically equivalent interpretation of $\psi\psi^*$) by Schrödinger.

If we do not integrate the equation (86 a) but substitute for the functions ψ_n their expressions (85 a), then this equation assumes the form:

$$\psi\psi^* = \sum_m \sum_n c_m^* c_n \psi_m^{0*} \psi_n^0 \, e^{i2\pi(\nu_m - \nu_n)t}. \tag{89}$$

It obviously contradicts the classical addition theorem for probabilities, according to which the probability ρ should reduce to the sum of the probabilities $\rho_{nn} = |c_n|^2 |\psi_n^0|^2$, for each of the alternative states taken separately. In addition, however, to these constant terms, the right side of (89) contains a number of terms of the type $\rho_{mn} = c_m^* c_n \psi_m^{0*} \psi_n^0 \, e^{i2\pi(\nu_m - \nu_n)t}$ which are due to a kind of *interference* between the vibrations of different frequencies and are themselves harmonic functions of the time. The sum of ρ_{mn} and ρ_{nm} is of course a real quantity which can be represented by the expression

$$a_{mn} \cos 2\pi\nu_{mn} t + b_{mn} \sin 2\pi\nu_{mn} t,$$

with real coefficients

$$a_{mn} = c_m^* c_n \psi_m^{0*} \psi_n^0 + c_m c_n^* \psi_m^0 \psi_n^{0*}, \qquad b_{mn} = i(c_m^* c_n \psi_m^{0*} \psi_n^0 - c_m c_n^* \psi_m^0 \psi_n^{0*}),$$

and

$$\nu_{mn} = \nu_m - \nu_n, \tag{89 a}$$

which denotes the oscillation frequency (with the sign + or −) of the

probability term, arising from the 'interference' of the two alternative states m and n. If we put in this formula

$$\nu_m = \frac{\epsilon_m}{h} = \frac{W_m + m_0 c^2}{h}, \qquad \nu_n = \frac{\epsilon_n}{h} = \frac{W_n + m_0 c^2}{h},$$

we get

$$\nu_{mn} = -\nu_{nm} = \frac{W_m - W_n}{h}. \tag{89 b}$$

We thus see that the *'interference freq uencies' ap pearingin* (89) *coincide with the 'transition freq uencies' of Bohr's theory.* These frequencies are the frequencies of the radiation, which according to this theory is emitted by the particle in a transition from the higher state (with greater energy) to the lower, or which is absorbed to effect the transition in the opposite direction. The particle we shall assume to be an electron.

It may be remarked that this important and very suggestive result is a consequence of the facts, (1) that the wave functions ψ_n were defined (to ensure the multiplication law of the probability theory) as *complex quantities* involving the time through the exponential factor $e^{-2\pi i \nu_n t}$ and not as real quantities involving the factors $\cos 2\pi\nu_n t$ and $\sin 2\pi\nu_n t$ in any other combination than $\cos 2\pi\nu_n t \pm i \sin 2\pi\nu_n t$, and (2) that the probability density has been defined as the product of ψ with the conjugate complex function ψ^*.

If, for instance, the wave functions were defined as the real parts of the above expressions ψ_n, and the probability density simply as the square of the corresponding wave functions, then the probability density ρ for a stationary state ψ_n would depend upon the time (namely, oscillate with a frequency $2\nu_n$), while in the general case of superposition of a number of such states we should get for ρ a sum of terms oscillating not only with the frequencies $\nu_m - \nu_n$ ('difference tones') but also with the frequencies $\nu_m + \nu_n$ ('summation tones'), which have no physical meaning.

The fact that, for a stationary state, ρ as defined above is independent of the time, whereas in the case of the superposition of, say, two stationary states ρ contains a term, oscillating with the 'optical' frequency associated with these two states, suggests the possibility of replacing Bohr's conception of the emission of radiation, corresponding to the quantum theory of radiation, by the ordinary radiation theory of classical electrodynamics. It has indeed been pointed out by Schrödinger that Bohr's assumptions, (1) that the stationary (quantized) states are non-radiative, and (2) that the optical frequencies multiplied

by h are equal to the energy difference between any two such states, can be deduced from classical electrodynamics if in the calculation of the electromagnetic field produced by the electron, the charge of the latter (e) *is assumed not to be concentrated at a definite point but to be spread through space with a volume density* $e\rho$.

In fact the distribution of charge in such an 'electronic cloud' in the case of a stationary state is independent of the time and must therefore give rise merely to an electrostatic field involving no radiation of energy.

If, on the other hand, two or more amplitude coefficients in (89), e.g. c_n and c_m are different from zero, then pulsations appear in the spatial distribution of the 'electron cloud' which, according to classical electrodynamics, produce an electromagnetic alternating field whose vibration frequency agrees with the frequency ν_{mn} of the pulsation.

On this mechanism the production of light vibrations is analogous to the production of sound vibrations through *beats* between two electrical high-frequency vibrations somewhat out of tune. This manner of electrically producing sound is used for many purposes in radio work (Whiddington's condenser, Theremin's radio music). On the probability theory of the wave-corpuscular parallelism the quantity $e\psi\psi^*$ should *not* be considered as the *true* charge density of a cloud-like electron (as was initially believed by Schrödinger), but only as the 'statistical' or 'probable' charge density, corresponding to a point-like electron, whose position is not exactly known. The effects depending on its position cannot be determined exactly; if, however, the probability of finding the electron in the volume-element dV is equal to $\psi\psi^* dV$, such effects will be the same *as if* the electron's charge were distributed with the volume density $e\psi\psi^*$.

From this point of view we could regard Schrödinger's electromagnetic field as the time-mean value of the 'true' field produced by the point-like electron according to the laws of classical electrodynamics. This interpretation does not seem, however, to have a physical meaning so long as the 'true' position of the electron cannot actually be determined. On the other hand, as has been shown at the beginning of this section, the superposition of the different stationary states implied by the formula $\psi = \sum c_n \psi_n$ must not be interpreted as the *coexistence* of these states, since they are 'alternative', that is, mutually exclusive.

We are thus led to the conclusion that Schrödinger's theory of light emission is not to be considered as a 'classical', that is, deterministic, theory but as a 'wave-mechanical' theory, involving the probability conception, through which it can and must be brought into (symbolic)

correspondence with the corpuscular or quantum theory of light emission due to Einstein and Bohr. The two theories of light emission do not contradict but rather correspond to each other in the sense of the wave-corpuscular parallelism. In the relation between light and matter, this parallelism is more complicated than for light alone (that is, the propagation and not the emission of light), or for matter alone (moving with constant energy). But its essential features remain the same as in these simple cases. The important point is not to mix corpuscular conceptions with wave conceptions, as was done in the classical theories of matter and radiation. If matter is to be described in corpuscular terms, i.e. if an electron is visualized as a point charge, then light must also be described in corpuscular terms and light emission visualized as the emission of a photon. If on the other hand we wish to describe light emission as the spreading of electromagnetic waves due to the electron's motion, then the electron must likewise be described as a wave system.

A new feature of the wave-corpuscular parallelism which emerges in connexion with light emission is the correspondence between the 'spontaneous transition' from a stationary state of higher energy to one of lower energy implied by Bohr's theory and the 'interference' between the electron waves associated with these two states implied by Schrödinger's theory.

The correspondence between waves and particles of a definite kind is characterized by two types of relations—the 'quality relations' $\epsilon = h\nu$, $\mathfrak{g} = h\mathbf{k}$, and the 'quantity relation' which asserts that the probability *of locating* the particle at a given point is $\psi\psi^* \, dV$.

In the same way the correspondence between the wave and the corpuscular theories of light emission can be characterized by relations of these two types. The quality relation is given in this case by Bohr's frequency condition $h\nu = W_m - W_n$, to which should be added an analogous equation between the momentum of the emitted light quantum and the recoil (i.e. change of momentum) experienced by the emitting system (hydrogen atom, or electron in its box). The 'quantity relation' must obviously connect the *probability of transition* (which now takes the place of the probability of location) with some quantity, quadratic in ψ, which specifies *the intensity of the light* emitted thanks to the 'interference' of the two states ψ_m and ψ_n between which the transition is supposed to take place.

According to the classical electrodynamics the electromagnetic energy emitted per unit time by an oscillating electron is equal to $2e^2w^2/3c^3$,

where w is the electron's acceleration. If the electron is supposed to oscillate parallel to the x-axis, $w = d^2x/dt^2$. For harmonic oscillations of frequency ν we have $x = a\cos 2\pi\nu t - b\sin 2\pi\nu t$ and $d^2x/dt^2 = -4\pi^2\nu^2 x$, so that the 'intensity' of the radiation, i.e. the average value I of the energy emitted per unit time is

$$I = \frac{2}{3c^3}(2\pi\nu)^4\pi\overline{x^2} = \frac{1}{3c^3}(2\pi\nu)^4(a^2+b^2). \tag{90}$$

The preceding results can easily be modified to fit the Schrödinger 'wave-mechanical' electrodynamics. The essential difference between Schrödinger's electrodynamics and the classical theory is that the coordinate x of the electron cannot be considered as a function of the time but must be estimated on the basis of the location probability $\psi\psi^*$. It seems natural therefore to replace x by its average or probable value

$$\bar{x} = \int x\psi\psi^*\, dV, \tag{91}$$

and apply the classical electrodynamics to the system *as if* x were the actual value of the electron's coordinate.

If ψ represents a stationary state $\psi = \psi_n$, then $\bar{x} =$ constant, $d^2\bar{x}/dt^2 = 0$, and there is no radiation. If on the other hand ψ represents the superposition of two such states,

$$\psi = c_m\psi_m + c_n\psi_n,$$

then
$$\bar{x} = c_m c_m^* x_{mm} + c_m^* c_n x_{mn} + c_m c_n^* x_{nm} + c_n c_n^* x_{nn},$$

where

$$\left.\begin{aligned} x_{mm} &= \int x\psi_m\psi_m^*\, dV = \text{const.},\\ x_{mn} &= \int x\psi_m^*\psi_n\, dV = \int x\psi_m^{0*}\psi_n^0\, dV\; e^{i2\pi(\nu_m-\nu_n)t} = x_{mn}^0\, e^{i2\pi\nu_{mn}t},\\ x_{nm} &= \int x\psi_m\psi_n^*\, dV = \int x\psi_m^0\psi_n^{0*}\, dV\; e^{i2\pi(\nu_n-\nu_m)t} = x_{nm}^0\, e^{i2\pi\nu_{nm}t},\\ x_{nn} &= \int x\psi_n\psi_n^*\, dV = \text{const.} \end{aligned}\right\} \tag{91 a}$$

Thus

$$\left.\begin{aligned} \frac{d^2\bar{x}}{dt^2} &= -(2\pi\nu_{mn})^2(c_m c_m^* x_{mn} + c_m c_n^* x_{nm}),\\ &= -(2\pi\nu_{mn})^2 \cdot 2[\mathrm{R}(c_m^* c_n x_{mn}^0)\cos 2\pi\nu_{mn}t + \\ &\qquad + \mathrm{I}(c_m^* c_n x_{mn}^0)\sin 2\pi\nu_{mn}t], \end{aligned}\right\} \tag{91 b}$$

where $\mathrm{R}(f)$ denotes the real and $\mathrm{I}(f)$ the imaginary part of f. The last equation follows from the fact that x_{mn} and x_{nm} are, by definition, conjugate complex quantities ($x_{mn}^* = x_{nm}$), as are also $c_m^* c_n$ and $c_m c_n^*$.

The preceding expression for $d^2\bar{x}/dt^2$ coincides with the classical one

if we put $\nu = \nu_{mn}$, $a = 2\mathrm{R}(c_m^* c_n x_{mn}^0)$, and $b = 2\mathrm{I}(c_m^* c_n x_{mn}^0)$. We are thus naturally led to replace the classical expression for the intensity of the radiation (90) by the following 'wave mechanical' expression:

$$I = 4\frac{(2\pi\nu_{mn})^4}{3c^3}|c_m|^2|c_n|^2|x_{mn}^0|^2. \tag{92}$$

If we now consider the process of light emission from the corpuscular point of view, the intensity of radiation must be defined as the product of the quantum of energy $h\nu$ emitted in the spontaneous transition from the upper of the two states, m say, to the lower n, by the probability that such a transition will actually take place in unit time. This probability must obviously be equal to the product of the probability $|c_m|^2$ that the electron is initially in the upper state, and of the probability A_{mn} that under this condition it will jump in unit time to the lower state. The coefficient A_{mn} is usually denoted as the 'transition probability $m \to n$'; it was introduced for the first time by Einstein in 1917. We thus get, according to the corpuscular picture,

$$I = h\nu_{mn}|c_m|^2 A_{mn}. \tag{92 a}$$

This expression differs from the expression (92) in the important point that it does not contain the factor $|c_n|^2$ giving the probability of the final state. The appearance of this factor in (92) actually excludes the possibility of the corpuscular (probability) interpretation of this expression unless we assume that the probability of the transition $m \to n$ is proportional not only to the probability of the initial (upper) but also to that of the final (lower) state. The fact that this assumption has no physical meaning is most clearly seen if we replace a single light-emitting electron by a large number N of such electrons, or rather atoms (which must be thought of as copies of the same atom). Then the products $N|c_m|^2$ and $N|c_n|^2$ can be interpreted as the number of atoms in the respective states. The number of atoms jumping from the state m to any other state n must be equal to a certain fraction A_{mn} of $N|c_m|^2$, which is *independent of the number of atoms in the other state*, and so the energy emitted in unit time corresponding to jumps $m \to n$ must be $h\nu_{mn} N|c_m|^2 A_{mn}$.

We thus see that the wave-mechanical expression (92) can be reconciled with the corpuscular one (92 a) only if we put $|c_n|^2 = 1$, which is inconsistent with the equation $\sum_n |c_n|^2 = 1$ and the condition that $|c_m|^2 > 0$.

This means that the wave-mechanical theory of radiation, outlined

above, requires some further modification. Whatever this modification may be, it is natural to assume that it will not fundamentally alter the expression (92) but will replace it by a similar expression not containing $|c_n|^2$, and identical with (92) for $|c_n|^2 = 1$.

With this assumption we can put

$$h\nu_{mn}|c_m|^2 A_{mn} = \frac{4(2\pi\nu_{mn})^4}{3c^3}|c_m|^2|x^0_{mn}|^2,$$

which leads to the following expression for the Einstein probability coefficient A:

$$A_{mn} = \frac{64\pi^4}{3c^3}\frac{\nu^3_{mn}}{h}|x^0_{mn}|^2. \tag{93}$$

In deriving this result we took into account only the x-component of the electron's motion. Taking account of the two other components, we get in the same way

$$A_{mn} = \frac{64\pi^4}{3c^3}\frac{\nu^3_{mn}}{h}(|x^0_{mn}|^2 + |y^0_{mn}|^2 + |z^0_{mn}|^2), \tag{93 a}$$

where y_{mn} and z_{mn} are harmonically oscillating quantities similar to x_{mn}, and y^0_{mn} and z^0_{mn} are their amplitudes similar to x^0_{mn}.

Quantities of the type x_{mn}, which can be defined for any function $f(x, y, z)$ by the formula

$$f_{mn} = \int f\psi^*_m\psi_n\, dV, \tag{94}$$

are called '*matrix elements*' of this function, because they form an ordered array or 'matrix':

$$\begin{vmatrix} f_{11} & f_{12} & f_{13} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ f_{21} & f_{22} & f_{23} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ f_{31} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{vmatrix}$$

The diagonal elements of this matrix refer to definite stationary states; they are constants which can be defined as *the average values* of the quantity f for these states. The non-diagonal elements are harmonic functions of the time

$$f_{mn} = f^0_{mn}e^{i2\pi\nu_{mn}t}, \tag{94 a}$$

with amplitudes given by

$$f^0_{mn} = \int f\psi^*_m\psi^0_n\, dV. \tag{94 b}$$

When $m = n$ the amplitudes coincide with the matrix elements themselves. The amplitudes can be interpreted from the corpuscular point

of view as determining the probabilities of transitions between the corresponding states under certain conditions specified by the character of the function $f(x, y, z)$. It should be remarked that the matrix elements of any *real* quantity (x, y, z) satisfy the so-called 'Hermitian' relation

$$f^*_{mn} = f_{nm}. \tag{94c}$$

The square of the modulus of f_{mn} can thus be written in the form

$$|f_{mn}|^2 = f_{mn} f^*_{mn} = f_{mn} f_{nm} = |f_{nm}|^2. \tag{94d}$$

We shall see in the next section that this symmetry relation can be considered as the general expression of the reversibility law, that is, of the equal probability of reverse transitions.

18. Wave-Mechanical Theory of Transition Processes

The wave-mechanical interpretation of spontaneous transitions, given in the preceding section, is not actually a theory of such transitions, but is rather an inference drawn in an indirect manner from the investigation of the electromagnetic field supposed to be produced by the system (atom, bound electron) in a state of motion associated with the wave function $\psi = \sum c_n \psi_n$. Remembering the meaning of the coefficients c_n, we can readily see that transitions from one state, m say, to another, n, must be connected with a change of the corresponding coefficients c_m and c_n, or rather their moduli, which determine the probabilities of these states (namely, with a decrease of $|c_m|$ and an increase of $|c_n|$, so that the sum $|c_m|^2 + |c_n|^2$ remains constant). An adequate wave-mechanical theory of transitions must therefore refer directly to the amplitude coefficients and consist of a system of equations determining their rate of change with the time t.

Before attempting to establish these equations we must make it clear under what conditions the coefficients c_n may be expected to vary with the time. It should be noted that only those transitions which are connected with a variation of these coefficients need be taken into consideration. It would be possible, namely, to imagine that in an assembly of N copies of the given system some copies undergo a transition $m \to n$, while an equal number of copies undergo the opposite transition, so that the relative numbers of copies in the corresponding states $N|c_m|^2$ and $N|c_n|^2$ remain constant. Such transitions must be left out of account, as pure fictions which are not correlated with any observable phenomena.

The transitions which we have considered in the preceding section are called 'spontaneous' because they are not due to the action of any

external forces other than those which determine the stationary states ψ_n. They express, therefore, a natural tendency of the system (atom) to drop from states of higher energy to those of lower energy, emitting the surplus energy in the form of electromagnetic radiation. This 'tendency' is found also in the classical electrodynamic theory where it is manifested in a *continuous* decrease of the mechanical energy of a bound electron whose motion generates spreading electromagnetic waves; the 'field energy' of these waves (distributed with the volume density $(E^2+H^2)/8\pi$) represents the mechanical energy, lost by the electron. Now classical electrodynamics provides a satisfactory explanation of this tendency, reducing it to the action of an additional force, which is not taken into account in the evaluation of the potential energy of the electron U, and which can be regarded as produced by the *electron itself through its own radiation field*. This 'radiative force' or 'force of radiative friction' is expressed by the formula (due to Planck)

$$\mathbf{F} = \frac{2e^2}{3c^3}\frac{d\mathbf{w}}{dt}, \tag{95}$$

where $\mathbf{w}$ is the electron's acceleration. It can easily be shown that the average work done by this force in unit time is equal to the average value of the energy radiated per unit time $2e^2\overline{w^2}/3c^3$ which was used above in connexion with Schrödinger's theory of radiation. In fact, multiplying (95) by the displacement of the electron in unit time, that is, by its velocity $\mathbf{v}$, and integrating with respect to t, we have

$$\int_0^t \mathbf{F}.\mathbf{v}\,dt = \frac{2}{3}\frac{e^2}{c^3}\int_0^t \mathbf{v}.\frac{d\mathbf{w}}{dt}\,dt = \frac{2e^2}{3c^3}\left\{[\mathbf{v}.\mathbf{w}]_0^t - \int_0^t w^2\,dt\right\}.$$

If the electron has an oscillatory motion (which it must always have if restricted to a finite region of space) the average value of the first term in the brackets $\{\}$ vanishes, so that we get

$$\overline{\mathbf{F}.\mathbf{v}} = -\frac{2e^3}{3c^3}\overline{w^2}.$$

The conception of the electron's action upon itself does not seem to have any physical meaning unless we imagine the electron to possess some extension in space, i.e. to have a finite volume within which its charge e is distributed. From this point of view a 'self-action' of the electron can be reduced to the mutual action of the spatially distinct elements of its electric charge. Applying this idea to the total electromagnetic field generated by an electron in accelerated motion, we get

in addition to the radiative force (95) a force of the type $-m\mathbf{w}$, which can be identified with the force of inertia if the coefficient m is defined as the *mass* of the electron. This electromagnetic theory of mass, developed by J. J. Thomson and H. A. Lorentz, actually amounts to a reduction of the laws of classical mechanics to those of classical electrodynamics with the additional principle that the *total* force acting on the electron—that is, the sum of the external force and of the self-force—must be equal to zero. At the same time the mechanical energy of the electron can be reduced to the 'field energy' of the electromagnetic field generated by it in the adjacent region of space (within say, a wave-length distance from the electron), while the energy localized at larger distances can be considered as radiation energy. As a matter of fact they become amalgamated into a single electromagnetic energy, satisfying the conservation law.

It would be out of place to develop these ideas at greater length here. The general conclusion which must be drawn from the preceding considerations is that the present wave-mechanical theory, dissociated from electromagnetic theory, should be regarded as a provisional approximation to the truth, and that in the future it will be possible to reduce it to a new 'quantum' electrodynamics in a way similar to that pointed out above in connexion with the classical theory. We shall come back to this question later on. For the present we are interested in one particular point only, namely, that the so-called 'spontaneous' transitions, associated with the emission of light, are actually *not* spontaneous, but are produced or 'induced' by the radiative self-force due to the electromagnetic field of the emitted light waves. We thus come to the conclusion that transitions between stationary states, specified by the functions ψ_n, can only occur as a result of the action of some additional or 'perturbing' forces, not incorporated in the potential energy function $U(x,y,z)$, on the basis of which these stationary states are determined. Such 'forced' transitions have indeed been visualized in Bohr's theory—especially those due to the action of light radiated from some external source, and we can feel satisfied with the result that the 'spontaneous' transitions need not be considered as an exception to this principle.

Now from the wave-mechanical point of view this principle obviously means that in the absence of perturbing forces (we need not be worried about the fact that they cannot be completely removed in reality) the coefficients $\dot{c}_n$ in the expression $\psi = \sum c_n \psi_n$ must remain constant. Our object will be to find a system of equations which will determine the

rate of change of these coefficients if 'perturbing forces' are introduced, which may be represented by the addition of a (relatively small) term U' to the original expression of the potential energy U. It may be remarked that the 'perturbation energy' U' can be an arbitrary function of the time, while U is a function of the coordinates alone.

The wave functions are defined by Schrödinger's equation

$$\nabla^2\psi_n + \frac{8\pi^2 m_0}{h^2}(W_n - U)\psi_n = 0$$

only so far as their amplitudes $\psi_n^0(x,y,z)$ are concerned. The time factors $e^{-i2\pi\nu_n t}$ ($\nu_n = W_n/h$) do *not* follow from this equation. It is, however, easy to modify it in such a way that it will determine ψ_n completely, i.e. both with respect to the coordinates and the time. This can be done in the simplest manner by using the fact that the product $W_n\psi_n = W_n\psi_n^0 e^{-i2\pi W_n t/h}$ is equal to $-\frac{h}{2\pi i}\frac{\partial}{\partial t}\psi_n$, and accordingly replacing the energy in Schrödinger's equation by the operator $-\frac{h}{2\pi i}\frac{\partial}{\partial t}$. The function $\psi_n = \psi_n^0 e^{-i2\pi W_n t/h}$ will then be a *particular solution* of the generalized equation

$$\nabla^2\psi + \frac{8\pi^2 m_0}{h^2}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial t} - U\right)\psi = 0, \tag{96}$$

which is satisfied equally by all the functions ψ_n since it does not contain the characteristic energy-value W_n. The general solution of equation (96) in virtue of its linearity will be a function of the type $\psi = \sum c_n\psi_n$ with arbitrary *constant* coefficients c_n.

It should be noticed that this equation can be applied to a potential energy function U *which contains the time explicitly*—in which case particular solutions of the type of harmonic oscillations corresponding to stationary states with a definite energy obviously do not exist.

Let us replace the function U in (96) by the sum $U+U'$. The resulting equation can be rewritten as follows:

$$\nabla^2\psi + i\frac{4\pi m_0}{h}\frac{\partial\psi}{\partial t} - \frac{8\pi^2 m_0}{h^2}(U+U')\psi = 0. \tag{96 a}$$

Let us now try to represent the solution of this equation in the same form $\psi = \sum c_n\psi_n$ as before. The ψ_n are thus to be particular solutions not of (96 a) but of (96), that is, representatives of stationary states corresponding to the potential energy U and not to $U+U'$. It will obviously be possible to satisfy (96 a) in this way if the coefficients c_n

are determined not as constant quantities but as certain functions of the time $c_n(t)$, taking arbitrary values c_n^0 for some initial instant $t = 0$.

In order to determine these functions we substitute the expression

$$\psi = \sum_n c_n(t)\psi_n \tag{96 b}$$

in (96 a), and take into account that with constant values of the c_n's it would satisfy the equation (96). We are thus left with

$$\sum_n \left(\frac{h}{2\pi i} \frac{dc_n}{dt} \psi_n + U' c_n \psi_n \right) = 0. \tag{97}$$

Now the product $U'\psi_n$ can be expanded in a series of the type

$$U'\psi_n = \sum_k U'_{kn}\psi_k, \tag{97 a}$$

with coefficients U'_{kn} which depend only on the time but not on the coordinates. In fact, on multiplying (97 a) by ψ_k^* and integrating over all space, we derive from the orthogonality property (87) of the functions ψ_k and the normality condition (85 b)

$$\int \psi_k^* U' \psi_n \, dV = U'_{kn}. \tag{97 b}$$

Thus the coefficients U'_{kn} are merely the *matrix elements of the perturbation energy* U', according to the general definition (94). If U' does not contain the time explicitly they reduce to harmonic functions of the time

$$U'_{mn} = U^0_{mn} e^{i2\pi\nu_{mn}t}.$$

Otherwise their dependence upon t is more complicated.

Substituting (97 a) in (97), we have

$$\frac{h}{2\pi i} \sum_n \psi_n \frac{dc_n}{dt} + \sum_n \sum_k \psi_k U'_{kn} c_n = 0.$$

On replacing the summation index n in the first sum by k and changing the order of summation in the second sum, this becomes

$$\sum_k \psi_k \left(\frac{h}{2\pi i} \frac{dc_k}{dt} + \sum_n U'_{kn} c_n \right) = 0.$$

If this equation is multiplied by ψ_m^* and integrated over all space, since the factor in brackets does not depend upon the coordinates, it reduces to

$$\frac{h}{2\pi i} \frac{dc_m}{dt} + \sum_n U'_{mn} c_n = 0 \tag{98}$$

(in virtue of the orthogonality and normality of the functions ψ_m). *This*

is precisely the system of equations, determining the rate of change of the coefficients c_n, which we have been looking for.†

In order to determine the probability of transitions from a certain state, ψ_k say, in which the system may be supposed to be at the initial instant $t = 0$, to other states, we must put $c_k^0 = 1$ and all the other $c_n^0 = 0$. The equations (98) reduce in this case approximately, for sufficiently small values of t, to

$$\frac{dc_m}{dt} \cong -\frac{2\pi i}{h} U'_{mk},$$

which gives, on integration,

$$c_m(t) = -\frac{2\pi i}{h} \int_0^t U'_{mk}\, dt \qquad (m \neq k). \tag{98 a}$$

We thus see that the initial rate of change of the coefficients c_m, corresponding to transitions $k \to m$, where k denotes the initial state of the system, is proportional to the matrix element U'_{mk} of the perturbation energy U'. Now since $c_m(0) = 0$, $|c_m(t)|^2$ measures the probability that the transition $k \to m$ has taken place during the interval of time from 0 to t. From (98 a) we see that this probability is proportional to the square of the modulus of $\int_0^t U'_{mk}(t)\, dt$. If the perturbation energy does not contain the time explicitly, we have

$$\int_0^t U'_{mk}\, dt = U'^0_{mk} \int_0^t e^{i2\pi\nu_{mk}t}\, dt = U'_{mk} \frac{e^{i2\pi\nu_{mk}t} - 1}{i2\pi\nu_{mk}},$$

so that

$$c_m(t) = -U'^0_{mk} \frac{e^{i2\pi\nu_{mk}t} - 1}{h\nu_{mk}}. \tag{99}$$

The time factor in this expression represents in general an oscillatory function, to which corresponds an oscillation of $c_m(t)$ about a small average value $\frac{U'^0_{mk}}{h\nu_{mk}} = \frac{U'^0_{mk}}{W_m - W_k}$ or of $|c_m(t)|^2$ about the average value

$$\overline{|c_m(t)|^2} = \frac{2|U'^0_{mk}|^2}{(W_m - W_k)^2}. \tag{99 a}$$

When, however, $W_m = W_k$, i.e. $\nu_{mk} = 0$, we have

$$c_m(t) = -\frac{2\pi i}{h} U'^0_{mk}\, t,$$

† These equations were derived by Dirac in 1926.

so that $c_m(t)$ steadily increases with time. The probability of the transition $k \to m$ referred to unit time is given in this case by

$$\frac{d}{dt}|c_m(t)|^2 = \frac{8\pi^2}{h^2}|U'^0_{mk}|^2 t, \tag{99 b}$$

whereas the average value of this probability is equal to zero when $W_m \neq W_k$. Thus transitions due to a perturbation not depending (explicitly) upon the time take place mainly between states of the same energy (i.e. belonging to a 'degenerate' set). Since the energy of a stationary state is a measure of the frequency of the associated vibrations, such states are said to be 'in resonance' with each other. It should be remembered that they must be alternative, i.e. satisfy the orthogonality relation $\int \psi_m^* \psi_n \, dV = 0$.

The possibility of transitions between states of different energy seems at first sight to contradict the principle of the conservation of energy. Such an inference is, however, wrong, for as a matter of fact the energy of the system is not defined exactly when we neglect the perturbation energy U' which is considered to be responsible for the transitions between states corresponding to the potential energy U alone. An exact description of the stationary states ought to be based on the total potential energy $U+U'$, and of course there would be no transitions between *these* states.

The transition probabilities are characterized by the 'non-diagonal' matrix elements of the perturbation energy. The 'diagonal' elements $U'_{nn} = U'^0_{nn}$ have a very simple meaning when the unperturbed system is not degenerate, that is, if different states ψ_n belong to different values of the energy W, or if the matrix elements of U' for different (degenerate) states belonging to the same value of the energy vanish so that transitions between them are impossible. There is then a one-to-one correspondence between the 'approximately' stationary states specified by the functions ψ_n and the actual stationary states belonging to the potential energy $U+U'$, U'_{nn} being approximately equal to the difference of the corresponding energy values:

$$\Delta W_n = U'_{nn} = \int U' \psi_n \psi_n^* \, dV. \tag{100}$$

This result is in agreement with a well-known theorem of classical mechanics according to which the additional value of the energy W due to a perturbation is equal, to a first approximation, to the average value of the perturbation energy U' for the corresponding unperturbed motion. A full discussion of this question, and incidentally a proof of

equation (100), will be given later on in Part II. From the 'symmetry' relation $|U'^0_{mk}|^2 = |U'^0_{km}|^2$ it follows that *the probability of the transition $k \to m$ is equal to that of the reverse transition $m \to k$.* This is the *law of reversibility* already alluded to and established in an exact form for transitions which can be described as a transmission through a given field of force (see § 15).

As another important illustration of the transition theory sketched in this section we shall consider the case of a perturbing force oscillating harmonically with the time. To be concrete let us consider the electric force $eE_0 \cos 2\pi\nu t$ acting on a bound electron in the x-direction, arising from monochromatic light waves of frequency ν sweeping through space. The amplitude of the electric field strength E_0 can be considered as a constant, not depending upon the coordinates of the electron, so long as the wave-length of the light is much larger than the 'size' of the atom, that is, the linear dimensions of the region in which the functions ψ_n are sensibly different from zero. Under this condition, which is actually fulfilled for ordinary light, the action of the light reduces to that of an oscillating electrostatic field.

The perturbation energy U' can be defined here by

$$U' = -xeE_0 \cos 2\pi\nu t = -\frac{exE_0}{2}(e^{i2\pi\nu t} + e^{-i2\pi\nu t}), \tag{101}$$

since $-\partial U'/\partial x$ is equal to the perturbing force. We thus get

$$U'_{mk} = -x_{mk}\, eE_0 \cos 2\pi\nu t,$$

or

$$U'_{mk} = -\tfrac{1}{2} ex^0_{mk}\, E_0 (e^{i2\pi(\nu_{mk}+\nu)t} + e^{i2\pi(\nu_{mk}-\nu)t}), \tag{101 a}$$

and consequently, according to (98 a),

$$c_m(t) = -\frac{ex^0_{mk}\, E_0}{2h}\left[\frac{e^{i2\pi(\nu_{mk}+\nu)t}-1}{\nu_{mk}+\nu} + \frac{e^{i2\pi(\nu_{mk}-\nu)t}-1}{\nu_{mk}-\nu}\right]. \tag{101 b}$$

An immediate consequence of these equations is that the probability of a 'forced' transition $k \to m$ produced by light is proportional to the probability of the corresponding spontaneous transition ($k \to m$ if $W_k > W_m$ or $m \to k$ if $W_k < W_m$), since both probabilities are proportional to the square of the modulus of x^0_{mk}. For consistency in the case of the spontaneous transition, only light due to vibrations in the x-direction should be taken into account. Another direct inference from (101 b) is that the probability of the forced transition $k \to m$ is the same as that of the reverse transition $m \to k$ (under the action of the same light). If $W_m > W_k$, then the first of these transitions means an absorption

of light and the second a forced or 'induced' emission of light. It is clear that the two probabilities have an appreciable magnitude only if one of the two 'resonance conditions',

$$\nu_{mk}+\nu = 0 \quad \text{or} \quad \nu_{mk}-\nu = 0,$$

is fulfilled, reducing (101 b) practically to

$$c_m(t) = -\frac{i\pi}{h} eE_0 x^0_{mk} t. \tag{101 c}$$

If $W_m > W_k$, i.e. $\nu_{mk} > 0$, then the first of the above conditions is out of the question and we must have $\nu = \nu_{mk}$, or, since $\nu_{mk} = (W_m - W_k)/h$,

$$W_m = W_k + h\nu,$$

which is Bohr's condition for the absorption of light. This absorption involves the transformation of a 'quantum' of light energy $h\nu$ into mechanical energy. In the opposite case we get an 'induced emission' of a light quantum with the condition

$$W_m = W_k - h\nu.$$

It should be remarked that our conclusion about the absorption or emission of light (in the form of one quantum) is only an indirect inference which ought to be substantiated by a direct calculation of the electromagnetic field associated with the motion under consideration. Further, it is interesting to note that the peculiar properties of light do not actually appear in the preceding analysis, since we have treated the light as an oscillating electrostatic field. If, therefore, we want to cling to the conception of light quanta or photons we must associate them not only with travelling electromagnetic waves, but also with an oscillating electrostatic field and indeed with a harmonically oscillating force of any other character. This conclusion seems to point to the result that light quanta must be considered as a pure fiction—inevitable, however, so long as we deal with matter from the corpuscular point of view (which after all may be a fiction too!).

Equation (101 c) shows that the probability of absorption of the light, that is, of an optical excitation of the atom, referred to unit time and measured by $d|c_m|^2/dt$ must be proportional to the time t. This result contradicts the experimental facts, according to which the relative number of atoms, supposed to be initially in the same state (e.g. the normal state), which are excited by light of a given intensity in unit time is proportional to the intensity but independent of the time. The theory, however, reproduces correctly the dependence of the transition probability on the intensity, since the intensity is proportional to E_0^2.

This discrepancy between experiment and wave-mechanical theory is explained by the fact that in reality light is never strictly monochromatic, as we have assumed it to be, but is a mixture of various frequencies extending over a small spectral region (the width of the spectral line). Thus the resonance condition $\nu_{mk} = \nu$ cannot be exactly fulfilled. In other words, the resonance between the transition frequency and the frequency of the light vibrations is never a *sharp* one. If the intensity of the light, as measured by the square of the amplitude of the electric vibrations, referred to the frequency range $d\nu$ is denoted with $E_\nu^2\, d\nu$, then the probability $|c_m(t)|^2$ of the atom getting to the (excited) state ψ_m in the time t will be expressed by the following formula

$$|c_m(t)|^2 = \frac{e^2|x_{mk}^0|^2}{4\hbar^2} \int \left| \frac{e^{i2\pi(\nu_{mk}-\nu)t} - 1}{\nu_{mk}-\nu} \right|^2 E_\nu^2\, d\nu,$$

which is obtained from (101 b) by dropping the first term in the brackets (which does not contribute sensibly to $c_m(t)$ for larger values of t), squaring, and replacing E_0^2 by $E_\nu^2\, d\nu$.

Now the factor

$$\left| \frac{e^{i2\pi(\nu_{mk}-\nu)t} - 1}{\nu_{mk}-\nu} \right|^2 = 4\, \frac{\sin^2\pi(\nu_{mk}-\nu)t}{(\nu_{mk}-\nu)^2}$$

in the integrand has, for not too small values of t, a very sharp maximum at $\nu = \nu_{mk}$, so that in performing the integration the other factor E_ν^2 can be replaced by its value for $\nu = \nu_{mk}$ and the integration over ν or, better, the difference $\nu - \nu_{mk}$ extended practically from $-\infty$ to ∞. Putting $\pi(\nu_{mn}-\nu)t = \xi$, we get

$$\int \frac{\sin^2\pi(\nu_{mk}-\nu)t}{(\nu_{mk}-\nu)^2}\, d\nu = \pi t \int_{-\infty}^{+\infty} \frac{\sin^2\xi}{\xi^2}\, d\xi = \pi^2 t,$$

so that

$$|c_m(t)|^2 = \frac{\pi^2}{\hbar^2} e^2 |x_{mk}^0|^2 E_{\nu_{mk}}^2\, t. \tag{102}$$

Dividing this expression by t we get the transition probability referred to unit time, which is, as it should be, independent of the time.

If instead of $E_\nu^2\, d\nu$ we introduce the average value of the volume density of the energy of the light vibrations in the spectral interval $d\nu$,

$$\rho_\nu\, d\nu = \frac{1}{2}\, \frac{(E_\nu^2 + H_\nu^2)}{8\pi}\, d\nu = \frac{E_\nu^2}{8\pi}\, d\nu,$$

we can express this probability Γ_{mk} in the form

$$\Gamma_{mk} = B_{mk}\, \rho_{\nu_{mk}}, \tag{102 a}$$

with $$B_{mk} = \frac{8\pi^3 e^2 |x^0_{mk}|^2}{h^2} = B_{km}. \qquad (102\,b)$$

The coefficient B_{km} refers to the reverse transition corresponding to the induced emission of light by a previously excited atom. These coefficients were also introduced by Einstein in 1917 together with the coefficient A_{mk} characterizing the probability of the spontaneous transition $m \to k$. Comparing (102 a) with the expression (93) for A_{mk}, we get the following relation:

$$\frac{A_{mk}}{B_{mk}} = \frac{8\pi\hbar\nu^3_{mk}}{3c^3} = \frac{8\pi\hbar}{3\lambda^3_{mk}}, \qquad (102\,c)$$

where λ_{mk} is the wave-length of the absorbed or emitted light. This relation was obtained by Einstein, without determining the coefficients A and B separately, from statistical considerations with the help of Planck's radiation formula (see Ch. VII). In the preceding theory we have assumed the electrical vibrations of the incident light to be parallel to the x-axis. For natural light without any definite direction either of polarization or of propagation, the average value of the energy referring to electrical vibrations in one particular direction must be equal to one-third of the total spectral density ρ_ν. Replacing $\rho_{\nu_{mk}}$ in (102 a), accordingly, by $\frac{1}{3}\rho$, we get for this case (isotropic radiation), to which Einstein's theory actually refers,

$$B_{mk} = \frac{8\pi^3 e^2 |x^0_{mk}|^2}{3\hbar^2}, \qquad (103)$$

and

$$\frac{A_{mk}}{B_{mk}} = \frac{8\pi\hbar\nu^3_{mk}}{c^3}. \qquad (103\,a)$$

The expression (93) for the coefficient A_{mk} was derived in the preceding section in an indirect manner—through a comparison with the classical theory of the 'probable' electromagnetic field generated by the electron. It seems possible to derive it in a way similar to that used above for the derivation of the coefficient B, that is, by considering the radiative 'self-force' acting on the electron as a perturbing force and defining the perturbation energy U' for the particular case of a motion parallel to the x-axis by the formula

$$U' = \frac{2e^2}{3c^3}\frac{d^3x}{dt^3},$$

which is obtained from (101) if the external force eE is replaced by the 'self-force' (95). We meet, however, at once with the difficulty of interpreting in wave mechanics the time derivative of x. We shall see later,

in Part II, that it is actually possible to find the corresponding wave-mechanical expression and we shall return, in Part III, to the question of radiative damping and 'spontaneous' transitions connected with this expression.

The 'transition probabilities' play in the new wave-mechanical theory just as important a role as the location 'probabilities' discussed before. The matrix elements f_{mn} of some function $f(x,y,z,t)$, representing the cause of the transition (that is, the perturbation energy), are related to the corresponding transition probabilities, which are proportional to $|f_{mn}|^2$, just as the wave functions ψ_n are related to the corresponding 'location' probabilities $|\psi_n|^2$.

The matrix elements f_{mn} (with properly chosen coefficients) could be defined accordingly as the 'amplitudes' of the transition probabilities. It is further plain that instead of the addition law of classical probability theory there must be an addition or 'superposition' law for the amplitudes of the transition probabilities referring to *alternative* transitions between the same states. This will involve the appearance of interference terms in the expression for the resulting probability (namely, the probability that the transition will be effected in any one of the alternative ways). This interference effect of the transition probabilities can be illustrated by the following example. Let us suppose that the transitions are due to two types of perturbing forces with the potential energy functions f and g. The 'probability amplitude' for a transition $m \to n$ produced by the combined action of these two forces will be equal, according to (94), to the sum $f_{mn}+g_{mn}$, where the first term obviously refers to a transition produced by the force f and the second to a transition produced by the force g. The mechanism of these transitions does not concern us; the only thing we need know about them is that they can be considered as alternative. Now the resulting transition probability (for any of the two possible mechanisms) will be proportional to the product

$$(f_{mn}+g_{mn})(f^*_{mn}+g^*_{mn}) = |f_{mn}|^2+|g_{mn}|^2+(f_{mn}g_{nm}+f_{nm}g_{mn}),$$

the last terms representing the interference effect alluded to. As a second example we shall consider a 'combined transition' taking place from m to n via any one of a series of intermediate states s. Let us assume, for the sake of generality, that the first step of this transition $m \to s$ is due to a perturbing function f and the second $s \to n$ to another perturbing factor g. These steps being independent, the probability amplitude for the resulting transition $m \to n$ with a *particular* s will be

equal to the product $f_{ms}g_{sn}$. Transitions $m \to n$ differing from each other through the intermediary state s must obviously be considered as alternative. Thus the resulting probability amplitude for a transition $m \to n$ under the combined action of f and g (the former coming in the first place and the latter in the second), which we shall denote by $(fg)_{mn}$, will be

$$(fg)_{mn} = \sum_s f_{ms} g_{sn}.$$

This equation can be derived from the definition (94) of the matrix elements (as the so-called law of 'matrix multiplication'). It can be conveniently illustrated by the phenomenon of the diffraction of cathode rays. If we imagine the incident electrons to be reflected from a series of plane lattices constituting a crystal, contact with any one of these lattices may constitute the intermediary state s, whereas m will represent the initial position of the electron, e.g. at the emitting cathode, and n the final position, e.g. on the fluorescent screen or the photographic plate. The interference pattern observed on the latter is due to the 'interference' of the different probability amplitudes in the expression $|(fg)_{mn}|^2$ for the resulting transition probability.

In conclusion a few words should be added as to the physical meaning of the transition processes, as they are defined and dealt with in wave-mechanical theory. Wave mechanics does not attempt to describe the course of these processes in time, in a way similar to the classical description, where all the intermediate steps of a transition appear as a continuous and perfectly determined series of exactly specified states. The time actually does not enter at all into the wave-mechanical description of a transition. *This does not mean*—as is often believed—*that the transition does not require any time, taking place as a sudden jump from one state to another.* It only means that the time element does not play any role in our description of transitions, so long as their course in time cannot be traced.

19. Vector Waves: Polarization of Light and Electron Spin

The hitherto neglected fact that light waves are characterized from the 'quality' point of view, not only by their frequency and wave-number vector $\mathbf{k}$, but also by their *polarization*, can be used to illustrate an important property of matter which emerges from comparing the corpuscular and the wave conceptions. In electromagnetic theory the polarization is determined by the direction of the electric vector of the wave, or the plane containing this vector and the direction of propaga-

tion. This definition supposes the plane of the electric vibration to remain fixed (plane polarized light). The main point which need concern us here is that for the same values of ν and $\mathbf{k}$ there exist *two* types of light waves, polarized in mutually perpendicular directions (planes), through the superposition of which we can represent any other elementary plane harmonic wave with the same $(\nu, \mathbf{k})$ and any type of polarization, plane or elliptical. We thus see that for the complete specification of an elementary light wave we must add to the above two parameters a third one, s say, capable of assuming *two values only*, associated with the two alternative directions—or 'states'—of polarization. These values can be fixed arbitrarily, as, for instance, $s = +1$ and $s = -1$. The number s may be called the 'polarization number', and may be dealt with on the same footing as ν and $\mathbf{k}$, with the important difference that its region of values is restricted to two points, whereas ν and $\mathbf{k}$ can assume an infinite number of different values. For standing light waves, enclosed in a box with perfectly reflecting walls, the quantities ν and $\mathbf{k}$—or the corresponding corpuscular quantities—energy $\epsilon = h\nu$ and momentum $\mathfrak{g} = h\mathbf{k}$ are determined by the three 'qualitative' quantum numbers n_x, n_y, n_z, assuming all integral values from 0 to ∞. To these we can add s as a *fourth quantum number* with the two possible values $+1$ and -1.

But this is not all. The three quantum numbers n_x, n_y, n_z—or the three projections of the momentum $\mathfrak{g}$ on the coordinate axes which they determine—are associated with the three coordinates x, y, z, which (along with the time t) are the independent variables of the corresponding wave functions $\psi_{n_x, n_y, n_z}(x, y, z)$. This association is, as we have seen above, not merely a formal one; it is a 'conjugation' implying the impossibility of a simultaneous determination of the two associated quantities. It seems therefore natural to add to these three coordinates a fourth one, ζ say, which shall correspond to s in the sense that x corresponds to g_x or to n_x. Now since s can take two values only, the new 'coordinate' must be also limited to two values, which can be fixed at random. We can take, for instance, the same two values $+1$ and -1 as in the case of s. It must not be believed, however, that there necessarily exists a correspondence between these *values* of ζ and the values of s. Just as x can take all possible values for a given value of g_x or n_x, so also can ζ take 'all possible values', i.e. both $+1$ and -1, for a fixed value of s (either $+1$ *or* -1).

The wave functions, describing not only the wave-length and the direction of the propagation of light but also its polarization, will thus

have the form

$$\psi_{n_x,n_y,n_z,s}(x,y,z,\zeta) \tag{104}$$

(the time being dropped).

The fourth coordinate ζ can be considered as equivalent here to a subscript serving to replace the *one* wave function $\psi(x,y,z)$ used heretofore by two different wave functions of the same type (i.e. two functions depending upon the usual coordinates x, y, z alone—and of course upon the time). These two wave functions may be denoted by

$$\left.\begin{aligned}\psi_{n,s,+1}(x,y,z) &= \psi_{n,s}(x,y,z,+1)\\ \psi_{n,s,-1}(x,y,z) &= \psi_{n,s}(x,y,z,-1)\end{aligned}\right\}, \tag{104 a}$$

where n is an abbreviation for n_x, n_y, n_z. They can be regarded as the two components of a certain vector quantity. As a matter of fact the light waves *are* vector waves, constituted by an oscillating electromagnetic field. The magnetic forces, which are relatively weak and unimportant so far as their action on electrons is concerned, may be disregarded. They correspond to a magnetic field numerically equal to and perpendicular to the electric field, and thus actually determined by the latter in conjunction with the direction of the vector $\mathbf{k}$. We can then describe light waves as waves of electric force (intensity) $\mathbf{E}$ and put $\psi_{+1} = E_x + iE_y$ and $\psi_{-1} = E_z$ (see below), since the ψ need not or even must not be real.

The corpuscular interpretation of the numbers s and ζ remains rather indefinite. It seems natural, however, to associate s with some vector defining the polarization and capable of assuming two directions, e.g. the directions of polarization, its magnitude being fixed. As to the number ζ, it must simply define the association of the photon with either one of the two functions ψ_{+1} and ψ_{-1}, which may denote two 'sub-states', alternative with respect to each other in the sense of the classical probability theory. From this point of view the probability of finding the photon in one of the two possible states of polarization in a volume element $dV = dxdydz$ will be proportional to the product of dV by a probability density defined by the equation

$$\rho = |\psi_{+1}|^2 + |\psi_{-1}|^2. \tag{104 b}$$

As a matter of fact, with the above definition of the functions ψ_{+1} and ψ_{-1} this sum reduces to $E_x^2 + E_y^2 + E_z^2 = E^2$, or rather $E_{0x}^2 + E_{0y}^2 + E_{0z}^2 = E_0^2$ (where E_0 is the amplitude of E), that is, to a quantity which actually measures the intensity of light or, in corpuscular language, the concentration of the photons. In other words, ρ is the probability per unit volume of finding a photon at the point in question.

Similar considerations may, and in fact must, be applied to cathode waves and electrons.

The analysis of the structure of atomic spectra on the basis of Bohr's theory has disclosed the fact that to each quantized state ('orbit') predicted by this theory and specified by three quantum numbers, n_x, n_y, n_z, there actually correspond *two* states with slightly different energies, forming what is called a 'doublet'. This doublet structure of Bohr's energy-levels is revealed not only by the number and position of emission or absorption lines, but—in a still more striking way—by the distribution of electrons between the different levels or states in complex atoms. It has been established semi-empirically that the deeper levels in such atoms—for instance, the K-levels or each level of the L-group, and so on—are always associated with *two* electrons. In other words—to use the language of Bohr's theory—each orbit, specified by three quantum numbers n_1, n_2, n_3, is represented—so far at least as the inner orbits are concerned—by *two* electrons.

This 'doubling' has been formally interpreted by W. Pauli by the introduction, in addition to the quantum numbers n_1, n_2, n_3 (which correspond to the three degrees of freedom usually attributed to an electron), of a fourth quantum number s, capable of assuming two different values only. The doublet states, referred to above, may obviously be described as corresponding to the same values of the ordinary quantum numbers n_1, n_2, n_3 and to different values of the additional quantum number s.†

Thus in the case of electrons we meet with a situation similar to that discussed above in connexion with photons.

The physical meaning of the fourth quantum number—in the sense of the corpuscular conception—was suggested in 1925 by Uhlenbeck and Goudsmit, who showed that the 'doubling phenomenon' could be explained quantitatively by assuming that an electron should be treated not as a point charge but as a small electrified sphere *spinning with a fixed angular velocity about an axis which can assume two opposite orientations* depending on the conditions which determine the character of the translational motion. Owing to its alleged 'spin' the electron must possess an angular momentum and a magnetic moment proportional to it, which have to be added (vectorially) to the corresponding quantities due to the revolution of the electron about the nucleus. In order to account for the experimental facts it is necessary to assume

† The connexion of this with the so-called 'Pauli's exclusion principle' will be discussed in the next chapter.

firstly that the *magnetic moment* due to the spin is equal to that due to the revolution of an electron in a one-quantum orbit (Bohr's magneton) and the *angular momentum* to one-half of the corresponding orbital value; and *secondly*, as has just been mentioned, that under given conditions only two opposite orientations of the spin (or magnetic axis) of the electron are possible, for instance, parallel and antiparallel to the axis of the orbit, or in the presence of a powerful magnetic field parallel and antiparallel to the direction of this field.

These two opposite orientations of the spin axis are obviously those features of the electron's motion which are specified by the two values of the additional quantum number s.

Turning to the wave-mechanical description of the electron's motion we can tentatively develop the analogy between the doubling phenomenon for electrons, that is, cathode waves, and the doubling phenomenon for photons corresponding to the polarization of the associated light waves. It may be imagined, accordingly, that the cathode waves (associated with the motion of a single electron) are specified by a 'vector'-function of the type (104) (with the quantum numbers n_1, n_2, n_3 standing for n_x, n_y, n_z), or by two scalar functions of the type (104 a) corresponding to the two values of the 'spin coordinate' ζ, which is the conjugate quantity to the 'spin quantum number' s.

This point of view has been developed by Pauli, and especially by Dirac who has succeeded in giving a full wave-mechanical interpretation of the 'spin' phenomena by allowing the numbers s and ζ to assume not two but four different values—in agreement with the requirements of the relativity theory. We shall develop this theory in detail later (Part II), and shall confine ourselves here to the following remarks.

(1) In the wave-mechanical theory (both of Pauli and Dirac) the electron is actually treated as a point charge, and not as a spinning sphere, the phenomena attributed to the spin resulting from the fact that the electron's motion is characterized not by one but by two (according to Pauli) or four (according to Dirac) wave functions, depending on the usual coordinates x, y, z, and of course on the time.

(2) The four values of the 'spin quantum number' in Dirac's theory actually represent a re-duplication of the usual two values ($s = +1$ and $s = -1$ say), which correspond to the opposite orientations of the electron's 'spin axis'. This re-duplication is due to the fact that in the relativistic wave-mechanical theory of the electron each set of the ordinary quantum numbers n_1, n_2, n_3 can be associated with two states of motion (irrespective of the spin doubling considered above) which

correspond *to opposite signs of the electron's rest mass in the expression of the energy*. In other words, the rest mass for one of the two states, the 'ordinary' one say, is positive and equal to m_0 as it should be, whereas for the other, which may be called the 'extraordinary' state, it turns out to be negative. This result corresponds to the ambiguity in the sign of the square root in the relativistic relation between the energy ϵ and momentum g of a free electron:

$$\epsilon = \pm c\sqrt{(m_0 c)^2 + g^2}.$$

This relation follows from the quadratic equation $\epsilon^2/c^2 - g^2 = m_0^2 c^2$, and can be rewritten in the form

$$\epsilon = \pm \frac{m_0 c^2}{\sqrt{(1 - v^2/c^2)}},$$

where v is the velocity of the electron.

In the corpuscular theory the negative sign is dropped on the ground that the corresponding expression 'has no physical meaning'. As a matter of fact an electron moving according to the relativistic mechanics and having at some initial instant a positive energy $\epsilon > 0$, will never, whatever the character of the forces acting on it, pass to an 'extraordinary' state with the energy $\epsilon < 0$. For the smallest value of the former $\epsilon = +m_0 c^2$ is separated from the largest value of the latter $\epsilon = -m_0 c^2$ (both of them corresponding to $v = 0$) by a gap $2m_0 c^2$ which cannot be bridged in classical mechanics. It is quite otherwise, however, in wave mechanics, for wave mechanics allows not only real but also *imaginary values* of the momentum g, and consequently of the velocity v. We recall that such imaginary values occur in the corpuscular interpretation of the phenomenon of the 'total reflection' of waves (cf. § 12). Thus there exists the possibility of bridging the gap between the ordinary and extraordinary states by a transition of the electron through intermediate states corresponding to imaginary values of its velocity. In fact, perturbing forces can actually induce transitions of this kind. It follows from this that in the relativistic wave mechanics the 'extraordinary' states must not be rejected but must be allowed just as well as the ordinary ones.

A possible physical meaning of these extraordinary states appears from the fact that a particle with a negative charge $-e$ and a negative mass $-m$ in a given electrical field must move in the same way as a particle with the positive charge $+e$ and a positive mass $+m$. Thus electrons in the extraordinary states should behave to a certain extent like protons—allowance being made for the difference of mass. This

result points to a possibility of reducing protons to 'extraordinary' electrons or rather to the local shortage of such extraordinary electrons in a world practically filled up with them. Let us imagine that if all the electrons were in extraordinary states they would fill the world so completely as to produce no resultant electric field. If now an extraordinary electron is removed from its place, the 'hole' left by it will produce the same electric field as a proton. Further, under the influence of an external electric field, the extraordinary electrons surrounding the hole—and in consequence the hole itself—will move like a proton, with an abnormally small mass. But it does not seem impossible according to Dirac, that this discrepancy will ultimately be explained by taking into account the mutual action of the electrons in the extraordinary states.

It may be remarked that so long as protons are regarded as real particles (and not 'holes') their motion can be described wave-mechanically in the same way as that of electrons. In this case, however, we must limit ourselves both with regard to electrons and protons to the non-relativistic description corresponding to Pauli's theory.

(3) The analogy between light and electron waves with respect to the doubling phenomenon proves to be very incomplete. The existence of polarization in the case of light waves is due to the transverse character of the electromagnetic vibrations constituting them. In the case of electrons the wave functions, $\psi(\zeta)$, cannot be associated with definite directions in space (or planes in the space-time extension of the relativity theory), so that it is impossible to speak about the direction of the vibrations defined by the $\psi(\zeta)$. This question will be examined in detail later (Part II).

We shall confine ourselves here to the remark that polarization of electrons has been investigated experimentally by a few physicists (Joffé and Rupp in particular) who have used for cathode rays the same type of apparatus with two rotating mirrors, as that used by Fresnel for the production and analysis of polarized light, but with an entirely negative result.

It must be admitted in general that the analogy between light and matter waves, which has led us so far, has a very limited significance and may become misleading if it is exaggerated.†

We shall return to this question in the next chapter from a somewhat different point of view.

† Just as this has happened to the analogy between light and sound, or elastic vibrations. This analogy greatly helped to establish the wave theory of light, despite the fact that the nature of the two phenomena was finally recognized to be entirely different.

IV

WAVE MECHANICS OF A SYSTEM OF PARTICLES

20. Method of Copies: Theory of Quantized Waves

In the wave mechanics of a single particle it is often very convenient to use, as we have already done, a kind of 'multiplex' or 'statistical' representation. Thus we imagine an 'assembly' or crowd made up of a very large number of *copies* of the particle and distributed in space or between the different states of motion in such a way that the relative number of copies in a given element of volume (dV) or in a given state (ψ_n) is proportional to the probability of finding one of them—i.e. the single particle which they represent—in this element of volume ($|\psi|^2\, dV$) or in that particular state ($|c_n|^2$). The absolute number of copies in the assembly is immaterial; we can imagine, for example, that the copies fill up the space in a continuous way (continuous assembly), their number being thus not only infinite but not even enumerable. The probabilities $|\psi|^2\, dV$ and $|c_n|^2$ are measured by the *relative* number of copies in the volume element dV or in the state ψ_n.

If, as for a bound particle, these probabilities can be normalized to 1 ($\int |\psi|^2\, dV = 1$, $\sum |c_n|^2 = 1$), then the total number of copies N can be taken as unity also.

This 'multiplex' representation of a particle can be carried out in two externally different but practically equivalent ways. It is possible namely to imagine all the copies to move in *the same* (*common*) *field* of force—including *a single copy of* the source of the latter, i.e. the positive nucleus in the case of a hydrogen atom or a box with reflecting walls in the case of an electron 'jailed' in such a box. Or we can imagine each copy of the particle to be associated with its own copy of the source of the field of force in which it is moving, these fields of force not overlapping one another (for different copies). The first picture permits of a statistical interpretation both of the 'location probability' $|\psi|^2\, dV$ and of the 'state probability' $|c_n|^2$; the second picture of the latter probability alone. The second picture is usually to be preferred if we are not interested in the location of the particle relative to the source of the field of force in which it is moving and are concerned only with the probabilities of the different quantized states or with the probabilities of transitions between these states. Our preference for this picture is partially explained by a tendency to consider different copies of the same particle not as 'ghosts' created by our own imagina-

tions but as real *specimens of this particle* which, as such, should act upon one another and so disturb the motion due to the external forces alone (supposed to be produced by a single common source). This tendency is stimulated by the fact that different specimens of an elementary particle—an electron, for example, or an atom of a definite kind—are actually as identical as if they were all copies of one particle. This identity of the elementary particles of which the world is constituted (apart from the duality between electrons and protons, which according to Dirac's view may be merely apparent) is to be regarded as a very fundamental and important fact of nature.

We are thus naturally led to the question, whether it is not possible to interpret the identity of all the elementary particles forming a given material system by treating them as copies of a single particle of the given sort. Of course the treatment sketched above should be modified in order to allow for the fact that these 'materialized ghosts' appear in a perfectly fixed number, N say, and that they exert a certain action on each other. But these are minor features in comparison with the chief (though less apparent) characteristic of the method under consideration, which consists in the fact that in treating the N particles as copies of a single one we are *stripping them of their individuality*, i.e. denying the possibility of discriminating between the 'different' particles or, in other words, denying the possibility of identifying 'one and the same' particle at different instants of time.

This circumstance becomes plain if we take into account the character of the information which we can expect to get (or to give) about our system of particles if we wish to treat them as copies of one of them. This information reduces to the probable *number of particles* in a given volume or in a given state. *Which* particles are to be found in this volume or in this state is a question that *cannot* be answered. In fact the whole system of particles is characterized by *one* set of coordinates x, y, z (and perhaps a fourth—spin or polarization—coordinate ζ) referring to *any* one of the particles. Before we proceed to apply the method of copies to a system of identical particles we must ask ourselves whether or not this complete 'collectivization' of such particles can be justified from the point of view of the wave mechanics of the separate particles.

Such a collectivization certainly *cannot* be justified from the point of view of classical mechanics. In the classical description of a system of N identical particles the latter are referred to as 'different' in the sense that it is possible—at least in principle—to distinguish 'which is

which'. This can be done by following their motion from a certain initial instant when they were labelled, for instance, by attaching to each of them a definite number of the series 1, 2,..., N, or by attaching a separate set of coordinates labelled with the corresponding number. Now this 'following' of the separate particles in their supposed motion is a thing which was *believed* to be possible so long as one did not question the validity of the fundamental principle of classical mechanics —the *principle of determinism*—as applied to an elementary particle. The preservation of their individuality by members of a crowd of similar particles is then a consequence of this principle. Wave mechanics does not adopt any such view. It now becomes impossible to follow the career of a particle or even to *think* of this career as a deterministic continuous motion. Even if the particle is associated with a wave packet we can neither locate it nor determine its velocity exactly. This uncertainty in terms of the classical states corresponds to an *effective* phase extension, equal (in the order of magnitude) at least to h^3. If therefore, instead of a single particle, we consider a system formed by N such particles, we must give up all attempts at their 'identification', or rather 'differentiation', and describe the state of the system exactly as if we had to do with utterly indistinguishable copies of a single particle.

This result, which is a direct consequence of wave mechanics as applied to a single particle, will be referred to henceforward as the 'identity principle'. It consists in recognizing the very simple and fundamental fact that particles identical in their nature cannot be distinguished from one another by their position or motion. These are actually indeterminate and specifiable in terms of probabilities only.

We shall become acquainted with the important implications of this fact when we come to study statistical problems. Turning back to the modifications which are required to apply the usual method of copies to a system of similar particles, we shall consider first only that modification which is necessary to express the fact that the number of particles (copies) has a *fixed integral* value (N), and we shall leave the mutual action between the particles out of account. In many cases this mutual action can be practically neglected, as, for instance, in the case of a beam of cathode rays, or it may be replaced approximately by the action of a suitably defined external field (see below).

This modification is possible because, in the wave mechanics of a single particle, the *absolute* value of the functions ψ_k representing the stationary states, or of the coefficients c_k in the compounded function

$\psi = \sum c_k \psi_k$ remains arbitrary. The obvious course to take is to remove this arbitrariness by introducing an additional condition for these quantities which should adequately express the fact that we are dealing with a definite number N of real particles. The simplest and most natural way to do this is to preserve the usual normalization condition for the wave functions ψ_k referring to the stationary states of an individual particle (we are considering here and in future the case of bound particles only) and to postulate that the quantities $|c_k|^2$ shall assume *integral values only*, representing the actual numbers of the particles in the corresponding states. We shall write this condition as follows:

$$|C_k|^2 = N_k = \text{whole number (including zero)} \tag{105}$$

(replacing small letters c by capital ones) where

$$\sum_k N_k = N \quad \text{(total number of particles).} \tag{105 a}$$

The function

$$\Psi = \sum_k C_k \psi_k \tag{105 b}$$

will accordingly be normalized to N, i.e. it must satisfy the equation $\int \Psi\Psi^* \, dV = N$, while the quantity $\Psi\Psi^* \, dV$ will measure the *probable number of particles* in the volume element dV and not the location probability for a single particle supposed to be in one of the alternative states ψ_k. This location probability will now be measured by the function

$$\psi = \frac{1}{\sqrt{N}} \Psi = \sum_k \frac{C_k}{\sqrt{N}} \psi_k = \sum_k c_k \psi_k \tag{105 c}$$

with amplitude coefficients c_k whose squares must bear to each other the ratios of the whole numbers N_k, their sum being equal to 1.

The above principle can be illustrated by the quantum theory description of radiation (e.g. temperature radiation) enclosed in a box with perfectly reflecting walls. From the corpuscular point of view this radiation is to be considered as an ideal gas consisting of a number of photons in different states of motion represented by the different quantized states. It should be remembered that this quantization is the result of (or rather is equivalent to) the nodal conditions for the standing light waves associated with the photons. This 'qualitative' quantization serves to determine the 'quality', i.e. the energy and momentum, of the 'different' photons. Actually they are as identical as the 'different' electrons in an electron gas (i.e. a system formed by a number of electrons whose interaction is either neglected or replaced by an equivalent external field), the 'qualitative' difference between them being due merely to the difference in the character of their motion.

Now besides this qualitative quantization another quantization process is needed because the number of photons associated with each of the quantized states must be a *whole* number or zero. This second quantization process is obviously equivalent to the condition (105) and will be denoted (for lack of a better term) as 'quantitative quantization'.

The term 'quantization' applied to the determination of the 'distribution' or 'partition' numbers N_k implies that these numbers should be regarded as a sort of new quantum numbers, similar in principle to the 'qualitative' quantum numbers n (or n_x, n_y, n_z) which serve for the specification of the states of individual particles. It is in fact a very attractive feature of the present theory (first presented in connexion with radiation by Debye in 1912 and further developed by Dirac and by Jordan) that it attempts completely to reduce the notion of the particle, or more exactly of all the particles of a given kind, to that of a system of waves—not, however, in the naïve sense of the original wave-packet theory of de Broglie, but in a much deeper sense, by applying the idea of quantization to the amplitudes of the various harmonic waves representing stationary states.

This theory of 'quantized waves' can be applied not only to light waves, but just as well to waves associated with the motion of identical particles of any kind. Such a theory must obviously treat the harmonic wave systems representing the different stationary states in a manner very similar to that inaugurated by Planck in his pioneer work on the quantum theory of the harmonic oscillator. According to Planck's theory a harmonic oscillator with a vibration frequency ν can have an energy equal to an integral multiple of the energy quantum $h\nu$. We can apply the same principle to each of the harmonic wave systems, representing a stationary state ψ_k by *treating this wave system as a quasi-material harmonic oscillator with a frequency* ν_k. A particle, when it is in a state described by this oscillator, has an energy $h\nu_k$. If there are N_k identical particles associated with this state the total energy is $N_k h\nu_k$. Now if this energy is ascribed not to the particles themselves, but to the harmonic oscillator representing their common state, then the number N_k assumes the role of the quantum number of the oscillator in question, and the fact that it has an integral value may then be considered as a consequence of the quantum theory.

It is true that according to the new quantum theory—or wave mechanics—the energy of a linear harmonic oscillator (with one degree of freedom) is *not* equal to an integral multiple of $h\nu$ but differs from it by a term $\frac{1}{2}h\nu$. This, however, does not seem to have much signi-

ficance. Admitting the value $(N_k+\frac{1}{2})h\nu_k$ for the energy E of a quantized wave, we can use the available energy, measuring the number of particles associated with this wave by $E/(h\nu)-\frac{1}{2}$.

The coefficients C_k in equation (105) are constant (complex) numbers. Putting $\psi_k = \psi_k^0(x,y,z)e^{-i2\pi\nu_k t}$ we can replace the product $C_k\psi_k$ by $\Gamma_k\psi_k^0$, where

$$\Gamma_k = C_k e^{-i2\pi\nu_k t} \tag{106}$$

is just that oscillating quantity which enables us to treat a harmonic wave system as a harmonic oscillator. We cannot, however, identify it with the coordinate Q_k of a linear oscillator, because it is a complex quantity while this coordinate must be real. The connexion between Γ_k and Q_k can be derived from the fact that the quantity $|C_k|^2 = \Gamma_k\Gamma_k^*$ multiplied by $h\nu_k$ must be equal to the energy of the oscillator expressed as a function of its coordinate Q_k and velocity dQ_k/dt or momentum $P_k = M_k\,dQ_k/dt$ (M_k = 'mass'). Thus

$$E_k = \tfrac{1}{2}\left[\frac{1}{M_k}P_k^2+(2\pi\nu_k)^2M_k\,Q_k^2\right] = \Gamma_k\Gamma_k^*h\nu_k,$$

which can be satisfied by putting

$$\Gamma_k = \frac{1}{\sqrt{(2h\nu_k)}}\left[\frac{1}{\sqrt{M_k}}P_k - i2\pi\nu_k\sqrt{M_k}\,Q_k\right]. \tag{106 a}$$

Denoting the real and imaginary parts of C_k by C_k' and C_k'', i.e. putting

$$C_k = C_k'+iC_k''$$

we get the following expressions:

$$P_k = \sqrt{(2h\nu_k M_k)}(C_k'\cos 2\pi\nu_k t + C_k''\sin 2\pi\nu_k t),$$

$$Q_k = \frac{\sqrt{(2h\nu_k M_k)}}{2\pi\nu_k M_k}(C_k'\sin 2\pi\nu_k t - C_k''\cos 2\pi\nu_k t),$$

which indeed satisfy the relation $P_k = M_k\,dQ_k/dt$, while Q_k represents the general solution of the equation of motion of a harmonic oscillator, i.e.

$$M_k\frac{d^2Q_k}{dt^2} = -\frac{\partial U_k}{\partial Q_k} \qquad (U_k = \tfrac{1}{2}(2\pi\nu_k)^2M_k\,Q_k^2).$$

In order to carry out the 'quantitative quantization' outlined above we must replace this 'classical' equation of motion by the corresponding quantum or wave-mechanical equation

$$\frac{\partial^2\Omega}{\partial Q_k^2}+\frac{8\pi^2M_k}{h^2}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial t}-\tfrac{1}{2}(2\pi\nu_k)^2M_k\,Q_k^2\right)\Omega = 0, \tag{107}$$

where Ω denotes a kind of wave function, determining in the usual way

the probability that Q_k will have a given value. (More exactly $\Omega\Omega^* dQ_k$ must be a measure of the probability that Q_k is contained in the interval dQ_k.) The particular solutions of this equation of the type

$$\Omega_{N_k} = \Omega^0_{N_k} e^{-i\frac{2\pi}{h}E_{N_k}t}, \tag{107 a}$$

satisfy the equation

$$\frac{d^2\Omega^0_{N_k}}{dQ_k^2} + \frac{8\pi^2 M_k}{h^2}(E_N - 2\pi^2\nu_k^2 M_k Q_k^2)\Omega^0_{N_k} = 0, \tag{107 b}$$

which determines the quantized values of the energy $E_N = (N_k + \frac{1}{2})h\nu_k$.

We thus see that, in carrying out the quantitative quantization, we must consider the coefficients Γ_k in the equation

$$\Psi = \sum_k \Gamma_k \psi^0_k(x, y, z) \tag{108}$$

not as functions of the time given by the relation (106), but rather as independent variables whose values can be specified in terms of probabilities only.† Thus they are similar to the spatial coordinates x, y, z or the coordinates Q_k.

We shall not go any further here into the mathematical theory of 'quantitative quantization', or, as it is often called, the theory of 'second' or 'double' quantization. It should be mentioned, however, that it must not be regarded as the only possible method of ensuring the 'whole number condition' for the quantities $|c_k|^2$. Indeed, we shall develop later (Part II) a somewhat different method giving similar results.

It must also be mentioned that these results need a substantial modification before they can be applied to electrons, protons, or any other particles which obey the so-called 'exclusion principle' of W. Pauli. This principle limits the possible values of the 'partition' (or 'quantum') numbers N_k to 0 and 1, whereas the method of quantization outlined above does not introduce an upper limit to N_k. We shall return to this question in § 23, where a detailed analysis of Pauli's exclusion principle will be given.

So far as the stationary states of a particle, or rather the standing harmonic wave systems associated with these states, can be represented by harmonic oscillators, a harmonic oscillator representing a state is to be regarded as something more fundamental than the particle occupying

† The derivative of Q_k with respect to the time t appearing in the definition of Γ according to (106 a) can be adequately replaced, as will be shown in Part II, by the differential operator $\frac{h}{2\pi i M_k}\frac{\partial}{\partial Q_k}$ acting on the 'wave-function' Ω.

this state. In particular, it may be possible for the state to be unoccupied; this means that the oscillator is in its lowest (normal) state with the quantum number N_k = zero. Thus the particles may be absent, but the oscillators are always there!

As a matter of fact, the notion of the particle becomes completely superfluous in this theory of 'quantized waves'. Only the *number* of particles appears, and even then merely as the quantum number of the oscillator representing a kind of a 'materialization' of these waves. We can, of course, assign to the wave a particle moving with an energy and momentum determined by the 'quality relations' $\epsilon = h\nu$, $\mathbf{g} = h\mathbf{k}$. But such a particle seems an artificial and useless appendage to the wave picture.

This reduction of the corpuscular conception to that of quantized waves or oscillators (the latter being, by the way, a corpuscular model of the former!) would appear quite satisfactory if the only thing we could observe about the particles was their number. As a matter of fact we can observe not only the number of the particles, but also—at least approximately—their position, motion, and interaction. It seems therefore improbable that the notion of particles could be altogether dispensed with. On the contrary the theory sketched above should be developed so as to reconcile completely the two aspects—wave and corpuscular—of the phenomena of nature.

So far as the mutual action of the particles is concerned, the generalization of the preceding theory does not present, in principle at least, any difficulties. The principle of this generalization can be illustrated by the theory of the interaction of the electrons emitted by an incandescent metallic filament in the neighbourhood of the filament, where the concentration of the electrons or the density of the resulting 'space charge' is considerable. The action of this space charge on an individual electron is equivalent to that of an 'external' electric field with a *given* potential distribution $\phi(x, y, z)$. In other words, the mutual action of the electrons is replaced by an *equivalent external electric field*, which, following Hartree, may be called the 'self-consistent' field. The potential of this field can be determined from Poisson's equation $\nabla^2\phi = -4\pi\rho$, where ρ denotes the density of the electric charge, i.e. the product of the charge of an electron e by the (average) number of electrons per unit volume at the corresponding point. Now in the wave-mechanical theory of a system of electrons, as developed in this section, based upon the method of copies, the probable number of electrons per unit volume is defined as the square of the modulus of the function Ψ,

We can therefore put $\rho = e\Psi'\Psi'^*$, and consequently

$$\nabla^2\phi = -4\pi e\Psi'\Psi'^*. \tag{109}$$

If the electric field determined by the potential ϕ were actually external in origin, then the potential energy of all the electrons with respect to it would be equal to the integral $\int \phi\rho\, dV$ or to the sum ϕe for all the N electrons. Since, however, it is produced by the electrons themselves, its energy is equal to half that sum. The corresponding potential energy referred to one electron is thus equal to

$$U' = \tfrac{1}{2}\phi e. \tag{109a}$$

If this energy is not too large, it can be considered as a *perturbation energy*, which may be supposed either to modify the stationary states ψ_k corresponding to the external potential energy $U(x, y, z)$ alone, or to cause *transitions* between these states. The modified stationary states can be determined by solving the equation

$$\nabla^2\Psi' + \frac{8\pi^2 m}{h^2}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial t} - U - U'\right)\Psi' = 0 \tag{109b}$$

in conjunction with (109) and (109a)—i.e. by simultaneously solving the 'equation of motion' (109b) and the equation for the space charge (109). We also need the normalization condition $\int \Psi'\Psi'^*\, dV = N$. The quantitative quantization can be undertaken after this solution has been obtained. If, on the other hand, we want to preserve the original definition of the stationary states and to regard the interaction energy U' as causing transitions between these states, then we must substitute the expression $\Psi' = \sum C_k\psi_k$ in equations (109) and (109b) and derive from these equations a system of equations for the functions $C_k(t)$. These equations will be cubic in the C_k (and not linear as in the case of an external perturbation). They can, however, be solved approximately without any quantization. The quantization of the quantities C_k (or $|C_k|^2$) can be achieved by considering the equations just mentioned as 'classical' equations for a system of harmonic oscillators whose motion is perturbed by coupling forces cubic with respect to their coordinates or momenta, and then replacing these classical equations by the corresponding quantum or wave-mechanical equations. Since, however, the oscillators representing different states of a particle cannot be treated as identical particles, the second (quantitative) quantization just indicated cannot be carried out by a repeated application of the method of copies outlined above.

As stated at the beginning of this section, the method of copies has

the advantage of interpreting, in what seems to be an adequate and natural manner, the fundamental fact that the world is made up of identical particles—or more exactly, of only three kinds of particles, electrons, protons, and photons. If it turns out to be possible to remove the duality of electricity by regarding the proton as being due to the absence of an electron in a world practically saturated with the latter, then we shall be left with the duality of electrons and photons alone. Assuming that the photons have just as real an existence as the electrons (and are not simply products of our imagination, due to an exaggeration of the analogy between cathode and light waves) it seems unreasonable to attempt to eliminate this remaining duality, for the photons are supposed to represent that electromagnetic field which is used for the description of the mutual action of the electrons (and protons). In so far as this mutual action is an essential feature of matter, the particles of matter must be treated in conjunction with the photons which carry the mutual action, as it were, from one particle (electron) to another.

It would thus be a great triumph of science if the world could be described as two connected systems of waves—cathode waves and light waves (or more generally the electromagnetic field)—with quantized amplitudes, the corresponding quantum numbers N_k being interpreted as the numbers of the associated particles. From this point of view the different processes taking place in the world or in any closed part of it could be interpreted as jumps in the amplitudes of the various wave systems representing the various quantized states of a number of electrons or photons. For instance, in the spontaneous emission of light by an electron jumping from a state m to a lower state n, the corresponding photon remaining within the enclosure containing the electrons, could be described by saying that the amplitude of the ψ_m-waves decreases by one quantum ($N_m \to N_m - 1$), while that of the ψ_n-waves simultaneously increases by one quantum. At the same time the amplitude of the light waves representing the emitted photon, and supposed to be included in the system of standing light waves representing the different states of a photon in the enclosure, also increases by one quantum.

This picture has, by the way, the advantage of allowing the interpretation of the *creation or annihilation of photons* in the processes of light emission or absorption, an interpretation which no other picture but this one—dealing with the numbers of photons in a given state as the vibrational quantum numbers of the associated waves—could afford. We might even say that the picture we are considering is 'too

good', for it would interpret the creation or annihilation of electrons just as easily as that of photons, whereas processes of this kind do not take place in reality. What probably does happen in reality (according to the weight of the evidence accumulated by astrophysics) is a *mutual annihilation of an electron and a proton* resulting in the appearance of *two* photons which carry away the energy and momentum. One photon would be insufficient to provide for the simultaneous fulfilment of the two conservation laws.† Probably the reverse process of the annihilation of two photons and the appearance at their expense of an electron and a proton also takes place. On the monistic theory of electrons and protons (due to Dirac) the mutual annihilation of an electron and a proton is to be considered as the transition of the electron into a state where its absence was interpreted as the presence of the proton. This transition takes place in two consecutive steps, each of which is accompanied by the emission and therefore the creation of a photon. Thus, according to this picture, the actual creation or annihilation of electrons, i.e. an increase or decrease in one of the quantum numbers N_1, N_2,... without a compensating change in the others, is impossible—a principle which, as we shall see later, can easily be incorporated in the quantitative theory based on the above picture.

21. The Method of Configuration Space

In spite of its theoretical advantages the theory of quantized waves proves to be unmanageable when we are dealing with a system of *different* particles, whether elementary, like electrons and protons, or quasi-elementary, i.e. treated as elementary particles, although in reality they may be clusters of elementary particles such as atomic nuclei, ions, atoms, and molecules. We shall limit ourselves provisionally to elementary particles which can be dealt with as *material points*. This means that, if they actually are clusters, their internal state, as described by the quantum theory, is assumed to remain invariable or to be specifiable by an 'inner' coordinate which has to be added to the three positional coordinates x, y, z.

The motion of the particle will be described accordingly by a wave function ψ depending on these three coordinates, and of course on the time t. The introduction of a fourth (inner) coordinate specifying the spin or the polarization, etc., does not substantially alter the following

† This is clear if we refer the motion of the electron and proton (before the act of mutual annihilation) to a system of coordinates with respect to which their centre of gravity remains at rest, the resulting momentum being thus zero, while the momentum of a separate photon must be different from zero.

considerations; in what follows this inner coordinate will be left out of account. Let us now consider the case of two *actually* different particles, e.g. an electron and a proton, with coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2) which, for the sake of brevity, will be denoted by x_1 and x_2 respectively, x standing for all three coordinates x, y, z. We shall imagine these particles to be held within a finite region of space by external forces and we shall neglect their mutual action. The motion of one of them can therefore be considered as *completely independent* of that of the others.

With this limitation the probability that at a certain time t the first particle will be found in a volume element $dV_1 = dx_1\,dy_1\,dz_1$ and the second in the volume element $dV_2 = dx_2\,dy_2\,dz_2$ of the same region V may be defined as the product of the corresponding probabilities for the separate particles. It is therefore equal to the product of $|\psi_1(x_1, t)|^2\,dV_1$ and $|\psi_2(x_2, t)|^2\,dV_2$, where ψ_1 and ψ_2 are the wave functions associated with the motion of the respective particles.

The validity of the classical multiplication law for the probabilities does not seem to require particular justification here. Yet if such justification is needed, it can be provided by the fact that this law has been found to be valid for the composition of three independent uni-dimensional motions in three mutually perpendicular directions into a single three-dimensional motion and shown to be applicable to the corresponding wave functions (probability amplitudes) themselves. In the latter case this multiplication law for the wave functions representing independent motions was restricted to the non-relativistic theory, implying the additivity of the energies of the constituent motions and of the associated frequencies.

In the present case this restriction is not necessary, and the wave function (probability amplitude) $\psi(x_1, x_2, t)$ can be defined quite generally and exactly as the product of the wave functions ψ_1 and ψ_2:

$$\psi(x_1, x_2, t) = \psi_1(x_1, t)\psi_2(x_2, t). \tag{110}$$

The physical meaning of this combined wave function is that the square of its modulus, i.e. the product $\psi\psi^*$, multiplied by the volume element $dV = dV_1\,dV_2 = dx_1\,dy_1\,dz_1\,dx_2\,dy_2\,dz_2$ of the six-dimensional 'configuration space' of the two particles gives the probability that they will be found in the configuration (or more exactly in the range of configurations) defined by this volume element. The term 'configuration' is an abbreviation for the expression 'positions of the two particles *at the same time*'. The time plays here an essential role only in the general

case in which the functions ψ_1 and ψ_2 represent the superposition of a few alternative states of each particle or a motion in a field of force which varies with the time. If ψ_1 and ψ_2 both refer to a single stationary state with energy $\epsilon_1 = h\nu_1$ for the first particle and energy $\epsilon_2 = h\nu_2$ for the second, then $\psi_1(x_1, t) = \psi_1^0(x_1)e^{-i2\pi\nu_1 t}$ and $\psi_2(x_2, t) = \psi_2^0(x_2)e^{-i2\pi\nu_2 t}$, and hence the product $|\psi_1|^2|\psi_2|^2$ has the same value independently of whether the time is chosen in the same way for both particles or not. Referring the positions of the particles to the same time, we get, according to (110),

$$\psi(x_1, x_2, t) = \psi^0(x_1, x_2)e^{-i2\pi\nu t}, \tag{110 a}$$

where $\psi^0(x_1, x_2) = \psi_1^0(x_1)\psi_2^0(x_2)$, $\nu = \nu_1 + \nu_2$. The latter equation corresponds to the additivity of the energies ϵ_1 and ϵ_2, which holds exactly both in the non-relativistic and the relativistic theory, so that the resulting frequency remains connected with the resulting energy by the usual relation $\epsilon = h\nu$. The interpretation of the 'probability amplitude' as a 'wave function', that is, as an expression describing the propagation of some sort of waves, can still be maintained in the case of the two-particle function $\psi(x_1, x_2, t)$ if we imagine the waves to travel in the six-dimensional configuration space $x_1, \ldots, z_2$. This is, of course, a mathematical fiction that has no physical meaning beyond the fact that the 'factorized' function (110) represents a combination of two independent three-dimensional wave systems. This fiction is, however, very convenient, and we shall use it later with due allowance for its rather limited physical meaning.

The preceding results can easily be generalized for a system consisting of an arbitrary number of particles. So long as the latter move independently, that is, exert no action on one another (they may of course be acted on by *external* forces), the wave function for the whole system can be represented as the product of the wave functions for the separate particles. This product function relates to a state of the whole system resulting from the combination of individual states of each particle, and the square of its modulus multiplied by a volume element of the configuration space gives the probability of the configuration defined by this element *for the combined state in question.*

Such combined states for which the dependence of the resulting ψ-function on the time is given by a factor $e^{-i2\pi\nu t}$ are *stationary* states of the system possessing a definite energy $\epsilon = h\nu$, and thus correspond to the quantized states of Bohr's theory. It is natural to assume that the superposition principle, i.e. the addition law for the probability amplitudes, holds just as well for the wave functions of a system of

particles as for individual particles. This result is to a certain extent implied in the multiplication law (110); for if ψ_1 and ψ_2 are themselves represented as the sums of several wave functions (relating to alternative states of the corresponding particles), then $\psi(x_1, x_2, t)$ will be equal to a sum of functions of the type (110a). It is important to notice that, in the general case, such a sum will not be expressible in the form of a product of two (or more) factors relating to the separate particles. We thus see that the independence of the particles, i.e. the absence of interaction between them, is expressed in general by the fact that the wave function ψ can be represented as a sum of terms each of which is the product of certain wave functions for the individual particles. We shall now find the differential equation which is satisfied by such a function, and shall then generalize this equation for the case when the particles exert a mutual action on one another.

For simplicity we shall begin by considering the case of *two* particles *without interaction* in a state of motion represented by the 'factorized' wave function (110).

The two factors will be assumed to satisfy the ordinary (non relativistic) Schrödinger equation, so that

$$\frac{\partial^2\psi_1}{\partial x_1^2}+\frac{\partial^2\psi_1}{\partial y_1^2}+\frac{\partial^2\psi_1}{\partial z_1^2}+\frac{8\pi^2 m_1}{h^2}\left[-\frac{h}{2\pi i}\frac{\partial}{\partial t}-U_1(x_1, y_1, z_1, t)\right]\psi_1 = 0,$$

and

$$\frac{\partial^2\psi_2}{\partial x_2^2}+\frac{\partial^2\psi_2}{\partial y_2^2}+\frac{\partial^2\psi_2}{\partial z_2^2}+\frac{8\pi^2 m_2}{h^2}\left[-\frac{h}{2\pi i}\frac{\partial}{\partial t}-U_2(x_2, y_2, z_2, t)\right]\psi_2 = 0,$$

where the functions $U_1(x_1, y_1, z_1, t)$ and $U_2(x_2, y_2, z_2, t)$ represent the potential energy due to the external forces and m_1, m_2 are the masses of the two particles (supposedto be different). Multiplying the first of these equations by ψ_2/m_1 and the second by ψ_1/m_2 and adding, we obtain a combined equation for the product function $\psi_1\psi_2 = \psi$, i.e.

$$\left\{\frac{1}{m_1}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial z_1^2}\right)+\frac{1}{m_2}\left(\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial z_2^2}\right)+\right.$$
$$\left.+\frac{8\pi^2}{h^2}\left[-\frac{h}{2\pi i}\frac{\partial}{\partial t}-(U_1+U_2)\right]\right\}\psi_1\psi_2 = 0.$$

This equation remains true for the sum of any number of such factorized functions. The generalization required to include the case in which the particles under consideration are subjected to an arbitrary mutual action determined by the *mutual* potential energy $U_{12}(x_2-x_1, y_2-y_1, z_2-z_1)$ is quite obvious: we simply add the term U_{12} to the sum of the potential

energies U_1+U_2 due to the external field of force. In this way we obtain the equation

$$\left[\frac{1}{m_1}\nabla_1^2+\frac{1}{m_2}\nabla_2^2+\frac{8\pi^2}{h^2}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial t}-U\right)\right]\psi=0, \tag{111}$$

where $U = U_1+U_2+U_{12}$ and

$$\nabla_k^2 = \frac{\partial^2}{\partial x_k^2}+\frac{\partial^2}{\partial y_k^2}+\frac{\partial^2}{\partial z_k^2}.$$

In the special case when U does not contain the time explicitly, this equation has particular solutions of the type $\psi = \psi^0 e^{-i2\pi Wt/h}$. These solutions represent the stationary states of the system, W being the total energy and ψ^0 satisfying the equation

$$\left[\frac{1}{m_1}\nabla_1^2+\frac{1}{m_2}\nabla_2^2+\frac{8\pi^2}{h^2}(W-U)\right]\psi^0=0. \tag{111a}$$

The rest energy $m_1c^2+m_2c^2$ can be left out of account. If U reduces to the sum U_1+U_2, ψ can be reduced to the factorized form considered above. The total energy W then splits up into a sum of two terms W_1 and W_2 representing the energies of the separate particles.

To illustrate equation (111a) we shall apply it to the simplest case of two particles *in the absence of external forces*. To make our problem more definite we shall consider a proton and an electron which together form a hydrogen atom ($U = U_{12} = -e^2/r_{12}$). We introduce the relative coordinates of the electron with regard to the proton $x = x_2-x_1$, $y = y_2-y_1$, $z = z_2-z_1$, and also the coordinates of the centre of mass of the atom

$$\xi = \frac{m_1x_1+m_2x_2}{m_1+m_2}, \qquad \eta = \frac{m_1y_1+m_2y_2}{m_1+m_2}, \qquad \zeta = \frac{m_1z_1+m_2z_2}{m_1+m_2}.$$

We thus obtain

$$\frac{\partial\psi}{\partial x_1} = \frac{\partial\psi}{\partial\xi}\frac{\partial\xi}{\partial x_1}+\frac{\partial\psi}{\partial x}\frac{\partial x}{\partial x_1} = \frac{m_1}{m_1+m_2}\frac{\partial\psi}{\partial\xi}-\frac{\partial\psi}{\partial x},$$

$$\frac{\partial^2\psi}{\partial x_1^2} = \frac{\partial^2\psi}{\partial\xi^2}\left(\frac{\partial\xi}{\partial x_1}\right)^2+2\frac{\partial^2\psi}{\partial x\partial\xi}\frac{\partial\xi}{\partial x_1}\frac{\partial x}{\partial x_1}+\frac{\partial^2\psi}{\partial x^2}\left(\frac{\partial x}{\partial x_1}\right)^2$$

$$= \left(\frac{m_1}{m_1+m_2}\right)^2\frac{\partial^2\psi}{\partial\xi^2}-\frac{2m_1}{m_1+m_2}\frac{\partial^2\psi}{\partial x\partial\xi}+\frac{\partial^2\psi}{\partial x^2},$$

and

$$\frac{\partial^2\psi}{\partial x_2^2} = \left(\frac{m_2}{m_1+m_2}\right)^2\frac{\partial^2\psi}{\partial\xi^2}+\frac{2m_2}{m_1+m_2}\frac{\partial^2\psi}{\partial x\partial\xi}+\left(\frac{\partial^2\psi}{\partial x^2}\right),$$

which gives

$$\frac{1}{m_1}\frac{\partial^2\psi}{\partial x_1^2}+\frac{1}{m_2}\frac{\partial^2\psi}{\partial x_2^2} = \left(\frac{1}{m_1+m_2}\right)\frac{\partial^2\psi}{\partial\xi^2}+\left(\frac{1}{m_1}+\frac{1}{m_2}\right)\frac{\partial^2\psi}{\partial x^2}.$$

We obtain similar formulae for the other two axes. Therefore, written in the new coordinates, equation (111 a) has the following form:

$$\frac{1}{\mu}\left(\frac{\partial^2\psi}{\partial\xi^2}+\frac{\partial^2\psi}{\partial\eta^2}+\frac{\partial^2\psi}{\partial\zeta^2}\right)+\frac{1}{m}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right)-\frac{8\pi^2}{h^2}\left(\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}+U\psi\right)=0, \quad (112)$$

where

$$\mu = m_1+m_2, \qquad \frac{1}{m}=\frac{1}{m_1}+\frac{1}{m_2}. \quad (112\,a)$$

Since, by our assumption, the potential energy U depends only on the relative coordinates, we can write the solution of (112) in the form

$$\psi = \psi_\mu(\xi,\eta,\zeta).\psi_m(x,y,z), \quad (113)$$

which is equivalent to treating the problem as one of two particles with masses μ and m and coordinates (ξ,η,ζ) and (x,y,z) respectively, which exert no action on one another. We obtain in this way, for the two factors in (113), the equations

$$\frac{\partial^2\psi_\mu}{\partial\xi^2}+\frac{\partial^2\psi_\mu}{\partial\eta^2}+\frac{\partial^2\psi_\mu}{\partial\zeta^2}-\frac{8\pi^2\mu}{h^2}\cdot\frac{h}{2\pi i}\frac{\partial\psi_\mu}{\partial t}=0 \quad (113\,a)$$

and

$$\frac{\partial^2\psi_m}{\partial x^2}+\frac{\partial^2\psi_m}{\partial y^2}+\frac{\partial^2\psi_m}{\partial z^2}-\frac{8\pi^2 m}{h^2}\left(\frac{h}{2\pi i}\frac{\partial}{\partial t}+U\right)\psi_m=0. \quad (113\,b)$$

The first determines the 'free motion' of the centre of mass of the given system (hydrogen atom), while the second determines the relative motion of one particle (electron) with respect to the other (proton). From the point of view of wave mechanics, this relative motion takes place exactly as if one particle (proton) were at rest, while the moving one (electron) had a mass $m=\frac{m_1 m_2}{m_1+m_2}$. This result completely agrees with that obtained from the usual corpuscular mechanics. Since the mutual interaction forces exerted on both particles must be equal and opposite to one another, we have

$$\frac{d^2x_2}{dt^2}=\frac{1}{m_2}F_x,\ldots; \qquad \frac{d^2x_1}{dt^2}=-\frac{1}{m_1}F_x,\ldots.$$

From this we obtain

$$\frac{d^2}{dt^2}(x_2-x_1)=\left(\frac{1}{m_1}+\frac{1}{m_2}\right)F_x,\ldots,$$

i.e.

$$m\frac{d^2x}{dt^2}=F_x=-\frac{\partial U}{\partial x},\ldots$$

in agreement with (113 b).

The preceding results can easily be generalized to an arbitrary number of different particles whose mutual action can be represented, according

to classical mechanics, by a certain potential energy function of their coordinates.

Such a method of treatment in which the corpuscular idea appears explicitly in the wave equation—each particle being represented by its coordinates—corresponds to Born's conception of the wave-corpuscular parallelism, a conception according to which only the corpuscles are regarded as being real, while the 'waves' in the multi-dimensional configuration space (defined by the coordinates of all the particles) merely determine the probability of certain corpuscular events.

All that has been said about the wave functions representing the motion of a single particle applies, with very little modification, to the wave functions representing, in the configuration space, the motion of a system of particles. In particular it is meaningless to talk of the absolute values of these wave functions. They can be normalized arbitrarily, for instance so that the integral, taken over the whole configuration space, becomes equal to 1 (provided, of course, that this integral converges, i.e. provided that the whole system of particles is 'bounded'). All the results referring to the stationary (quantized) states of a single particle, to transitions from one stationary state to another produced by perturbing forces, etc., can be applied to a system of particles if *this system is considered as a single particle in the configuration space.* We could even go so far as to use the method of copies discussed in the preceding section to represent the motion of a number of identical systems of particles. We shall not, however, penetrate into this rather arid region, but shall, on the contrary, apply the method of the configuration space to the special case of a system of identical particles.

22. The Symmetry Properties of Functions representing a System of Identical Particles

In applying the general theory to this special case we must introduce an important amendment, the object of which is to remove that individuality of the 'different' particles which is implied in assigning to each of them a separate set of coordinates. To incorporate the 'identity principle' into the method of the configuration space, we must limit ourselves to such wave functions $\psi(x_1, x_2, \ldots, x_N, t)$ as give rise to a probability function $\psi\psi^*$, independently of the way in which the 'different' particles are labelled. In other words, the probability function must be invariant with respect to any permutation of the indices 1, 2,..., N, or, to put it more simply, it must be *symmetrical with respect to all* the coordinate triplets (x_1, y_1, z_1), (x_2, y_2, z_2),..., (x_N, y_N, z_N).

This 'symmetry condition' for $\psi\psi^*$ is of course satisfied if ψ is itself symmetrical in all the particles. The latter condition, though sufficient, is by no means necessary. Another very simple and general condition consistent with the symmetry of $\psi\psi^*$ is that ψ should be an *antisymmetrical* function, i.e. alter its sign when any two particles are interchanged, which implies a 'transposition' of the corresponding coordinate triplets. The general condition for ψ is that *any* permutation of these coordinate triplets should result in ψ being multiplied by a factor of the form $e^{i\alpha}$, where α is some real quantity depending upon the character of the permutation—for such a factor will not affect the value of $\psi\psi^*$. (We shall see that this condition reduces to one of the two alternative conditions stated above.)

For the symmetry condition for $\psi\psi^*$ to have a physical meaning it must be *permanent*; i.e. if the symmetry be postulated for some particular instant, then it must be maintained for all time as a consequence of the wave-mechanical equation of motion. This equation, which is an obvious generalization of (111), assumes, in the special case of N identical particles, the following form:

$$\frac{h}{2\pi i}\frac{\partial\psi}{\partial t} = \left[\frac{h^2}{8\pi^2 m}(\nabla_1^2+\nabla_2^2+\ldots+\nabla_N^2)-U\right]\psi. \tag{114}$$

The potential energy $U(x_1, x_2, \ldots, x_N, t)$, in view of the identity of the particles, must be a symmetrical function of all the coordinate triplets. If, therefore, ψ is a symmetrical or an antisymmetrical function of these coordinate triplets, the product $U\psi$ must also be symmetrical or antisymmetrical. The same is true of the function resulting from ψ by the application of the symmetrical differential operation $\nabla_1^2+\nabla_2^2+\ldots+\nabla_N^2$. We thus see that if the function ψ is symmetrical or antisymmetrical at the instant t_0, it will be so at the next instant, i.e. when

$$\psi = \psi_{t_0} + \left(\frac{\partial\psi}{\partial t}\right)_{t_0} dt$$

and consequently for all time. This result is easily generalized to functions with any other symmetry character consistent with the symmetry of the product $\psi\psi^*$, the symmetry character of ψ being determined by the numbers α_P in the factors $e^{i\alpha_P}$ by which ψ is multiplied when a permutation P is applied to the particles. This permanence of the symmetry character of any function is actually more than is required by the identity principle, i.e. by the permanence of the symmetry condition for the product $\psi\psi^*$. It shows that the particles of a given kind must

be described by a wave function of some perfectly definite symmetry character which must be fixed once and for all. In the choice of this function we cannot be guided by the general principles of wave mechanics. These leave us free to make any choice we like. For the solution of this question we must therefore turn to experimental evidence. Before we consider this evidence we shall discuss the preceding results from a somewhat different point of view which will help in the elucidation of the theory, and will make it easier to compare the theory with experiment.

In a theoretical description of a system of identical particles, it is usually convenient to assign to each of them an individual state of motion, their mutual action being either neglected or accounted for approximately by the method of the 'self-consistent field' explained in § 20.

Such a description is used, for instance, in the case of a system of electrons moving about the positive nucleus of an atom or ion. According to the old quantum theory developed by Bohr, each electron was supposed to move in its own individual quantized orbit. The action of the other electrons upon it was accounted for approximately by introducing a certain 'screening constant' or 'effective nuclear charge', which had different values for electrons moving in different orbits.

If the individual states of motion ('orbits') are indicated by 1, 2,..., s, then the 'aggregate state' of the whole system can be described by specifying the distribution of the particles among these states, i.e. by giving the number of particles N_1, N_2,..., N_s associated with each of them ($N_1+N_2+\ldots+N_s = N$). These numbers will be referred to as 'partition numbers'.

We have already used this type of description in § 20, defining each individual state of motion k by a certain wave function $\psi_k(x, y, z, t)$ with a single set of coordinates for all the particles. Replacing the method of copies by the method of the configuration space, we must assign to each particle its own set of coordinates, and specify the particles occupying each state *as if* they were different. Now the assignment of particles to states can be carried out in a variety of different ways. One of them is to suppose that the first N_1 particles (labelled as 1, 2,..., N_1) are in the first state, the next N_2 (i.e. from N_1+1 to N_1+N_2) in the second state, and so on. The wave function ϕ representing the corresponding aggregate state of the whole system will be equal to the product of the functions ψ_k representing the individual states, the arguments being determined by the above assignment. Writing x_l for

all the coordinates of the lth particle, we thus have

$$\phi = \psi_1(x_1)...\psi_1(x_{N_1})\,\psi_2(x_{N_1+1})...\,\psi_2(x_{N_1+N_2})...\,\psi_s(x_{N_1+...+N_{s-1}+1})...\,\psi_s(x_N). \tag{115}$$

The interchange of particles belonging to the same state will not affect this function; if, however, we apply a permutation

$$P = \begin{pmatrix} 1, & 2, ..., & N \\ p_1, p_2, ..., & & p_N \end{pmatrix}$$

which involves interchange of particles belonging to different states, a new function of the type ϕ is obtained, which may be denoted by $P\phi$ or ϕ_P. The total number of such functions is determined by the partition numbers N_1, N_2,... according to the formula

$$g = \frac{N!}{N_1!\,N_2!...N_s!}. \tag{115a}$$

We shall assume henceforward that the individual states refer to the *same* (effective) external field. When applied to a system of electrons moving about a positive nucleus, this assumption means that the effective charge of the nucleus is the same for all the electrons. This would be approximately true if the nuclear charge were much larger than the total charge of all the electrons concerned. Hence the different functions ψ_k can be regarded as *orthogonal to one another* (they would not be orthogonal if they referred to motions in different external fields). It can easily be shown that, under this condition, the different functions ϕ_P must also be orthogonal to one another. In fact the integral $\int \phi^*\phi_P\,dV$ taken over the whole configuration space ($dV = dV_1\,dV_2...dV_N$) is equal to the product of integrals of the type $\int \psi_k^*(x_l)\psi_{p_k^{-1}}(x_l)\,dV_l$, where p_k^{-1} is the number derived from k by the permutation P^{-1} which is the reciprocal of P. Since at least some of the numbers p_k^{-1} are different from k, some of these integrals must vanish, and consequently $\int \phi^*\phi_P\,dV = 0$ or, more generally,

$$\int \phi_P^*\phi_Q\,dV = 0 \qquad (\phi_P \neq \phi_Q), \tag{115b}$$

which means that ϕ_P and ϕ_Q represent alternative aggregate states of the system.

All these states have the same energy,

$$E = N_1W_1 + N_2W_2 + ... + N_sW_s, \tag{115c}$$

where W_k denotes the energy of a particle in the individual state ψ_k (the mutual potential energy of the particles U' being neglected), and therefore they form a *degenerate set*. Degeneracy of this kind,

corresponding to the interchange of similar particles between different individual states, is called 'interchange' or 'exchange' degeneracy.

The g functions ϕ_P do not satisfy the identity principle. However, by means of the superposition process, they can be used to construct a set of functions

$$\chi = \sum c_P \phi_P \tag{116}$$

which satisfy this principle, i.e. which possess symmetry properties expressed by equations of the type

$$Q\chi = e^{i\alpha_Q}\chi, \tag{116a}$$

where $Q\chi = \chi_Q$ denotes the function obtained from χ by applying the permutation Q to the arguments $x_1, \ldots, x_N$.

These equations, together with the normality condition $\int \chi\chi^* \, dV = 1$ which reduces to

$$\sum |c_P|^2 = 1 \tag{116b}$$

because of the mutual orthogonality of the functions ϕ_P, serve to determine the coefficients c_P.

Applying the permutation Q to both sides of equation (116), we have, according to (116a),

$$\sum c_P Q\,P\phi = e^{i\alpha_Q} \sum c_P\, P\phi.$$

The 'product' QP of the permutations Q and P means that they must be carried out successively in the order indicated, reading from right to left. Putting $QP = R$, we have $P = Q^{-1}R$, where Q^{-1} is the permutation reciprocal to Q, so that QQ^{-1} is equal to the 'identical permutation'

$$\begin{pmatrix} 1, & 2, \ldots, & N \\ 1, & 2, \ldots, & N \end{pmatrix}$$

(denoted usually as 1). Further, since $\sum c_P\, P\phi = \sum c_R\, R\phi$ (the summation in both cases being extended over all the g effective permutations), we have

$$\sum c_{Q^{-1}R}\, R\phi = e^{i\alpha_Q} \sum c_R\, R\phi,$$

whence it follows that

$$c_{Q^{-1}R} = c_R\, e^{i\alpha_Q}. \tag{117}$$

The fact that the coefficients c'_R and c_R in an equation of the form $\sum c'_R \phi_R = \sum c_R \phi_R$ must be equal follows from the mutual orthogonality of the functions ϕ_R. In fact, multiplying this equation by ϕ_S^* and integrating over the whole configuration space, we get $c'_S = c_S$, the integrals involving all the other coefficients vanishing.

The relation (117) shows that all the coefficients c_P in (116) have the same modulus and differ from one another only with regard to what may be called their 'phase'. Since $\sum |c_P|^2 = 1$, we get $|c_P| = 1/\sqrt{g}$. Applying the above relation to the special case when R denotes the

'identical permutation', and putting for this case $c_R = |c_R| = 1/\sqrt{g}$ (which does not involve a loss of generality, for χ can be multiplied with any factor of the type $e^{i\alpha}$), we get

$$c_{Q^{-1}} = \frac{1}{\sqrt{g}} e^{i\alpha_Q}.$$

If we replace Q^{-1} by P and take into account that $\alpha_{P^{-1}} = -\alpha_P$, we get

$$c_P = \frac{1}{\sqrt{g}} e^{-i\alpha_P}. \tag{117 a}$$

The relation $\alpha_{P^{-1}} = -\alpha_P$ or $\alpha_{Q^{-1}} = -\alpha_Q$ follows from equation (116 a) by applying the permutation Q^{-1} to both sides. Applying some other permutation P to (116 a) we get

$$PQ\chi = e^{i\alpha_Q} P\chi = e^{i(\alpha_Q+\alpha_P)}\chi.$$

But by definition $$PQ\chi = e^{i\alpha_{PQ}}\chi.$$

Therefore $$\alpha_{PQ} = \alpha_P + \alpha_Q + 2\pi n, \tag{118}$$

where n is a whole number. Thus to the multiplication (i.e. successive application) of permutations there corresponds an addition of the associated numbers α (referring to the same χ-function of course), and eventually the addition of a whole number which is irrelevant so far as the factors $e^{i\alpha}$ are concerned.

We thus see that the correspondence between the permutations P and the numbers α_P is not a one-to-one correspondence, but that the same number α_P is, in general, associated with many different permutations. All these permutations (R) can be derived from one of them, P, by means of the formula

$$R = QPQ^{-1}, \tag{118 a}$$

where Q is an arbitrary permutation. In fact, since $\alpha_Q = -\alpha_{Q^{-1}}$, we have, according to (118), $\alpha_R = \alpha_Q + \alpha_P + \alpha_{Q^{-1}} + 2\pi n = \alpha_P + 2\pi n$.

Permutations of the type (118 a) form what is called a *class*, specified by any one of them. We can thus say that the numbers α_P are functions of the class to which the corresponding permutations belong, so that we can write

$$\alpha_P \equiv \alpha_R \,(\mathrm{mod}\ 2\pi). \tag{118 b}$$

The simplest class of permutations is that formed by *transpositions*, i.e. by permutations involving the interchange of *two* elements only. A transposition T is obviously equal (equivalent) to its reciprocal T^{-1}, whence it follows that $\alpha_T \equiv -\alpha_T \,(\mathrm{mod}\ 2\pi)$, that is, $\alpha_T \equiv 0$ or $\alpha_T \equiv \pi$. This means that the function χ must be either symmetrical or antisymmetrical with regard to different pairs of the arguments (electrons).

Now according to (118 b) it cannot be symmetrical with regard to *some* pairs and antisymmetrical with regard to others, but must be either symmetrical or antisymmetrical with regard to all of them. We thus see that the only functions ψ (or χ) allowed by the identity principle are the symmetrical and the antisymmetrical ones.

Since any permutation Q can be represented as the product of a certain—even or odd—number of transpositions, it must give when applied to an antisymmetrical function either $\alpha_Q = 0$ for an 'even permutation' or $\alpha_Q = \pi$ for an 'odd permutation'. For a symmetrical function we must, of course, always have $\alpha_Q = 0$.

It should be mentioned that if we did not restrict ourselves to functions satisfying the identity principle (i.e. to the symmetry condition for the product $\psi\psi^*$), we could form with the help of equation (116) a number (g) of linearly independent functions χ such that the result of the application of *any* permutation P to one of them is equal to a linear combination both of this function and of the others, according to the equation

$$P\chi_k = \sum_l P_{kl}\chi_l \qquad (k, l = 1, 2, \ldots, g).$$

The coefficients P_{kl} corresponding to each permutation P form what is called a 'representation' of this permutation.

The functions χ can be defined so as to consist of a certain number s of separate sets, characterized by the fact that the result of the application of any permutation to a function of a given set is equal to a linear combination of the functions of the same set only. The smallest sets obtainable by subdivision of suitably defined χ-functions form the basis of the so-called 'irreducible representations' of the permutations P, one such representation (of all the $n!$ permutations) corresponding to each of these sets.

From this point of view the symmetrical and the antisymmetrical functions χ are distinguished by the fact that they each form a separate set, consisting of one element only.

The product $\chi_k \chi_k^*$ for a certain function of a given set is not invariant with regard to all the permutations; it may, however, be invariant with regard to some of them. It can be easily shown, however, that the sum $\sum \chi_k \chi_k^*$ taken over *all* the functions of the same set, enjoys this invariant property; it is thus equivalent in this respect to the product $\psi\psi^*$ for the symmetrical or the antisymmetrical functions.

The different sets of functions χ associated with given values of the partition numbers $N_1, N_2, \ldots, N_s$, can be arranged in a definite order, corresponding, for example, to the increase of the number of functions

in each set. At the head of this series we shall always find the symmetrical function which can be formed with any values of the partition numbers. The next set, which is obtained if there is at least one individual state different from the rest, consists of N functions, which form the basis of the *same* representation of all the permutations, irrespective of the particular values of the partition numbers. Such sets of functions χ corresponding to different values of the partition numbers or in general to different systems of individual states ψ_k, can be treated accordingly as *equivalent* with each other.

As the number g of alternative aggregate states increases, the number of different (non-equivalent) sets of χ-functions increases also, new sets, corresponding to new irreducible representations of all the permutations, being added to the old ones. This series terminates with an antisymmetrical function which is obtained only if all the individual states are different.

We shall not stop here to study this question in greater detail. It presents a purely mathematical interest, being intimately connected with group theory. So long, however, as we restrict ourselves to the consideration of those functions only which satisfy the identity principle (i.e. the symmetry condition for $\psi\psi^*$ or $\chi\chi^*$) we need not enter into these mathematical intricacies and can leave all other functions χ out of account as devoid of physical meaning.

Taking for the sake of illustration a system of two particles we must consider two cases, corresponding to $s = 1$ and $s = 2$. Putting $s = 1$, we get only one function ϕ, namely,

$$\phi = \psi_k(x_1)\psi_k(x_2),$$

which is symmetrical with respect to x_1, x_2 and can therefore be identified with the single function $\chi = \chi_1$.

If instead of one individual state ψ_k we consider *two* different individual states ψ_k, ψ_l, then we get two different functions of the type ϕ, namely,

$$\phi_1 = \psi_k(x_1)\psi_l(x_2), \qquad \phi_2 = \psi_k(x_2)\psi_l(x_1),$$

from which two functions χ can be constructed, namely, the symmetrical function

$$\chi_1 = \frac{1}{\sqrt{2}}(\phi_1+\phi_2) = \frac{1}{\sqrt{2}}[\psi_k(x_1)\psi_l(x_2)+\psi_k(x_2)\psi_l(x_1)] \tag{119}$$

and the antisymmetrical function

$$\chi_2 = \frac{1}{\sqrt{2}}(\phi_1-\phi_2) = \frac{1}{\sqrt{2}}[\psi_k(x_1)\psi_l(x_2)-\psi_k(x_2)\psi_l(x_1)]. \tag{119a}$$

Thus in the case $s = 2$, we have the same symmetrical χ-function as in the case $s = 1$, and in addition to it an antisymmetrical χ-function.

These results can easily be generalized to any number of particles N. If all the particles are supposed to be in the same individual state ψ_k, we get only one χ-function, namely, the symmetrical function

$$\chi = \phi = \psi_k(x_1)\psi_k(x_2)...\psi_k(x_N). \tag{120}$$

As the number of individual states (s) increases we get in addition to the symmetrical function (which is formed by adding the g different factorized ϕ-functions and dividing the result by $\sqrt{g}$) a number of other sets of χ-functions which have no physical meaning. Finally with N different individual states $\psi_1, \psi_2, ..., \psi_N$, each of which is associated with one particle, we get $N!$ different factorized functions ϕ_P and a number of sets of χ-functions extending from the symmetrical function

$$\chi_1 = \frac{1}{\sqrt{(N!)}} \sum_P \phi_P \tag{120 a}$$

to the antisymmetrical. The latter is obtained from the general formula $\chi = \sum c_P \phi_P$ by putting $c_P = \pm 1/\sqrt{N!}$, the upper sign corresponding to an *even* permutation, i.e. an even number of transpositions (interchanges of any two x_l), and the lower sign corresponding to an odd permutation. The resulting function can therefore be written in the form of a determinant:

$$\chi_{N!} = \frac{1}{\sqrt{(N!)}} \begin{vmatrix} \psi_1(x_1) & \psi_1(x_2) & \cdot & \cdot & \cdot & \psi_1(x_N) \\ \psi_2(x_1) & \psi_2(x_2) & \cdot & \cdot & \cdot & \psi_2(x_N) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \psi_N(x_1) & \psi_N(x_2) & \cdot & \cdot & \cdot & \psi_N(x_N) \end{vmatrix}. \tag{120 b}$$

If two or more of the individual states coincide, i.e. if two or more particles are associated with the same individual state, this determinant vanishes (since two or more of its rows become identical). A motion described by an antisymmetrical function implies therefore the distribution of all the particles between different states.

It should be remembered that the functions (120 a) and (120 b) do not represent *exact* solutions of the equation (116), since they are expressed by means of individual wave functions $\psi_k(x)$ which can be defined only if the mutual action of all the particles is neglected or replaced by a 'self-consistent' external field which is but approximately equivalent to this interaction. The exact solution of the equation (116) can be, however, represented as a *superposition* (i.e. sum) of functions

of the type (120 a) or (120 b) with coefficients which must be defined as certain functions of the time. Since these coefficients must be interpreted as the probability amplitudes of the corresponding states, their variation with the time means that the mutual action of the particles, considered as a perturbation, will cause transitions of the whole system of particles from one aggregate state to another. The important point to note is that this state will *always* remain either symmetrical or antisymmetrical, a transition from one of them to another being impossible. This result can be proved directly by considering the matrix element of the perturbation energy U' with regard to two such states, χ_s and χ_a say. Since U' is a symmetrical function of all the variables x, as well as χ_s, while χ_a is an antisymmetrical function, the integral $\int U' \chi_s^* \chi_a \, dV$ which determines the probability amplitude of the transition $\chi_s \to \chi_a$ must vanish.

It follows in particular that if all the particles are distributed initially among different individual states, their motion being described by an antisymmetrical function, they will always remain distributed between different individual states.

23. The Pauli Exclusion Principle and its Consequences

We are now in a position to compare the theory with experimental data and to determine the symmetry character of those functions χ by which particles of a given kind must be described.

Direct evidence of this kind for electrons is revealed by a study of the structure of the more complex atoms. We need not recapitulate the chemical, spectroscopic, and X-ray data which led Bohr in 1921 to formulate his theory of atomic structure, according to which the electrons are distributed about the nucleus in 'groups' or 'shells' corresponding to successive values of the principal quantum number (n). Within each group there are sub-groups corresponding to successive values of the angular quantum number (l). A remarkable feature of this picture—a feature which Bohr's theory did not attempt to explain but merely postulated empirically—was that the number of electrons associated with the same type of motion (i.e. moving in similar orbits) could not exceed 2.

This limitation was remarkable because it contradicted the hitherto accepted principle that the normal state of an atom or ion must be the state of lowest energy. Now it can easily be shown that this lowest aggregate state—at least if the nucleus has a sufficiently high positive charge—would be realized if all the (N) electrons were in the lowest

individual state, in other words, if they were all moving about the nucleus in the smallest 'one-quantum' orbits. In reality (as revealed, for example, by the analysis of X-ray spectra) only two electrons are found in this state (corresponding to the K absorption limit). Furthermore, the next lowest state ($n = 2$, $l = 0$) is also occupied by two electrons only, and so on until we come to the 'external' electrons which determine the chemical and optical properties of the atom and which are found sometimes in pairs and sometimes alone in each state.

It should be remembered that, in Bohr's theory, each (individual) state was specified by three quantum numbers—the principal quantum number n, the angular quantum number l, and the axial quantum number m (cf. § 14). The reason why the maximum number of electrons which could be associated with the same state was 2, and not 1 or 3, was not clear. It became somewhat clearer when it was discovered that Bohr's stationary states were actually *double* and when this doubling was interpreted in terms of the electron's spin (§ 19). It then became possible to assert that *each individual state can be associated with one electron only*. This principle was formulated by W. Pauli in 1925 and was shown later (by Fermi, Pauli himself, and Sommerfeld) to apply not only to bound electrons in atoms and ions, but also to the 'electron gas' supposed to be formed by the 'free' (conductivity) electrons in a metallic body.

Pauli's principle is usually known as the 'exclusion principle', the term 'exclusion' indicating that a second electron cannot be admitted into a state already occupied by an electron. A distribution of electrons (or of any other similar particles) among any set of individual states satisfying this principle will be called an *exclusive distribution*. In such a distribution the partition numbers N_1, N_2,..., N_s have the values 1 or 0 according as the corresponding state is occupied or not.

Now in the preceding section we have seen that there exists a wave function χ which vanishes identically when the distribution is not exclusive, i.e. when one at least of the partition numbers is larger than 1. This is the *antisymmetrical* function expressed by the determinant (120 b). Pauli's exclusion principle can thus be incorporated in the wave-mechanical theory of a system of electrons if it be assumed that *such a system must be described by an antisymmetrical function*. This illuminating interpretation of Pauli's principle is due to Dirac.

It should be mentioned that the 'antisymmetry principle' of Dirac is not exactly equivalent to Pauli's principle in its original formulation as a restriction to exclusive distributions. For such distributions are

consistent not only with an antisymmetrical function but are also consistent with $N!-1$ other functions of higher symmetry, including the symmetrical function (120 a). We thus see that Dirac's antisymmetry principle is more restrictive than that of Pauli. However, they become equivalent if Pauli's principle be considered not from a purely systematic, but from a *dynamic* point of view, i.e. if it be required that a system of electrons supposed to be originally in an (aggregate) state satisfying Pauli's principle should not be able to pass, in the course of time, into another state violating this principle. Such transitions, destroying the exclusive character of the original distribution, would inevitably occur if the system of electrons were described by any function χ other than the antisymmetrical function.

We can thus say that Dirac's antisymmetry principle is the condition which provides for the continued maintenance of Pauli's principle throughout all time.

Another important point arises in connection with the description of a system of identical particles given in the preceding section. We assumed that each particle is specified by *three* coordinates x, y, z, to which corresponds the specification of the individual stationary states by three quantum numbers, in agreement with Bohr's theory. But to account for the doubling phenomenon, i.e. the splitting of each of the above states into two (associated with opposite orientations of the spin), we must add to the usual quantum numbers (n_x, n_y, n_z) a fourth 'spin quantum number' s capable of assuming only two values ($s = \pm 1$). Because of this we must *add to the three spatial coordinates* x, y, z a fourth 'spin coordinate' ζ, also taking only two values, $+1$ and -1 (cf. § 19). This modification does not, however, affect the results of the preceding section about the functions χ—or in particular the expression (120 b) for the antisymmetric function—if it be understood that x_l is an abbreviation for the *four* coordinates x_l, y_l, z_l, ζ_l. Of course the general proof of the permanence of the symmetry character of a multi-dimensional function $\psi(x_1,\ldots,x_N)$, which was given at the beginning of § 22, on the basis of the wave equation (114) not involving the spin, ought to be given on a broader basis of a generalized equation of this type, taking account of the doubling phenomenon. We shall return to this question in Part II after we have established such an equation. It will also be necessary then to revise the theory of transition-probabilities.

Leaving these questions aside, we notice that a function $\chi(x_1, x_2,\ldots, x_N)$ of the groups of four coordinates actually represents a set of 2^N functions

of the corresponding spatial coordinate triplets. This is especially clear if χ is expressed in terms of the individual functions $\psi(x_l, y_l, z_l, \zeta_l)$, since each of the latter is equivalent to two functions of the spatial coordinates (referring to the same stationary states), the spin coordinate playing the role of an index (cf. § 19).

If the doubling (or spin) phenomenon is disregarded, Pauli's principle can be stated as follows: *distributions for which the partition numbers are larger than* 2 *are excluded*. Using for the definition of stationary states the ordinary wave functions of the spatial coordinates alone, we must introduce, for the description of a system of N electrons, besides the antisymmetrical function, a number of functions χ with a higher symmetry, which do not vanish for the allowed distributions and which vanish identically for the excluded ones.† Each (allowed) distribution will thus be represented by a number of different states of (approximately) the same energy, forming what is called a 'spin multiplet'. Transitions between the different states of the same multiplet and, in general, between any two states which are described by χ-functions with a different symmetry character will be impossible provided we neglect the magnetic interaction due to the spin of the electrons. (We shall see later that such intercombination can occur if the 'spin interaction' is taken into account.)

It is interesting to notice that the permanence of the exclusive distribution follows from classical mechanics provided that the interaction between the particles is neglected and only external forces are considered. According to Liouville's theorem, to which we have already referred in § 11, the volume of the phase-space assigned to a particle, whose initial state has been specified inexactly, remains constant, i.e. independent of the time. If, instead of a single particle, we imagine a number (N) of such particles distributed over the 'uncertainty region' of the phase-space and moving according to the laws of classical mechanics in the same *external* field of force without interaction, just as if they were copies of the same particle, then Liouville's theorem can be formulated in the following way, equivalent to the preceding one: the concentration of the particles in the phase-space, i.e. their number per unit volume of this space, must not change with the time.

† In the case of two electrons, for example, the function (119) symmetrical with respect to the two coordinate *triples* is allowed by Pauli's principle just as well as the antisymmetrical one (119 a). The former must be multiplied by an antisymmetrical function of the two spin coordinates ζ_1, ζ_2, and the latter by a symmetrical function of these coordinates; this makes them both antisymmetrical with respect to the two coordinate quadruples (cf. Part II).

Now an individual state in the sense of wave mechanics implies a phase extension of volume h^3 in the sense of classical mechanics. If we replace the 'different' electrons by copies of the same electron and the wave-mechanical states by the corresponding regions of the classical phase-space, then the exclusive distribution can be defined as that distribution for which the concentration of the 'copies' in the phase-space is nowhere larger than 2 per h^3 (the spin being disregarded). If, instead of wave mechanics, we apply to these 'copies' the laws of classical mechanics, then it follows from Liouville's theorem that the concentration of the copies will be not only *nowhere* but also *never* larger than 2 per h^3, i.e. that the distribution will remain exclusive. This result can be generalized to include interaction between the particles so far as this interaction can be described by a self-consistent field.

We have talked until now of electrons only. But some direct evidence favours the natural assumption that Pauli's exclusion principle, or rather Dirac's antisymmetry principle, applies equally to *protons*.† This evidence, which is given by a study of the band spectrum emitted (or absorbed) by a hydrogen molecule and of the heat capacity of the hydrogen gas, will be examined in detail in Part III. There is also a great weight of indirect evidence. It is derived from data (mainly on band spectra) referring to more complicated particles, such as atomic nuclei, which are built up of several protons and electrons.

The symmetry character of the wave functions which are to be associated with such complex particles (treated as material points) can be inferred from the antisymmetry principle for electrons and protons by means of the following argument.

Let us suppose that we have to deal with a system of similar particles, formed by the combination of n electrons and m protons (the particles may be atomic nuclei, ions, atoms, or molecules). Interchange of the coordinates of any two such particles in the wave function χ describing their motion is equivalent to the simultaneous interchange of the coordinates of n electrons and m protons in another function X of the coordinates of these elementary particles (and not of the centres of gravity of the clusters formed by them) describing their motion in a more complete way. Since the interchange of any two electrons or any two protons must change the sign of the function X because of its supposedly antisymmetrical character, it will finally appear multiplied

† Protons possess a spin of the same magnitude as electrons so far as the mechanical moment of momentum is concerned. The associated magnetic moment, being inversely proportional to the mass, is 1,845 times smaller for a proton than for an electron.

by $(-1)^{m+n}$. This result must obviously remain true for the approximate or, better, incomplete function χ.

We thus arrive at the conclusion that a system of identical complex particles, considered as material points, must be described either by symmetrical or antisymmetrical functions depending upon whether these complex particles consist of an even or an odd number of elementary particles (electrons and protons). All *neutral* particles—such as atoms and molecules—must, in particular, be described by symmetrical functions. This result may be applied to photons if they are regarded as neutral particles. It must be admitted that this extension of the symmetry principle to particles which, if they really exist, are certainly not complex, does not appear to be well justified. A better justification, however, is provided by a comparison with the wave theory of light which does not impose any restriction whatever on the energy associated with a given system of standing light waves, or, in other words, on the number of photons associated with a given stationary state (represented by these waves). In particular, it is possible for all the N photons forming the system under consideration to be associated with the same individual state. The only function χ which does not vanish identically for this distribution is the symmetrical function. It must therefore represent the aggregate state of a system of photons for a distribution of any kind.

A further confirmation of this conclusion will be obtained later in connexion with the theory of temperature radiation.

The relation between the two methods of treating a system of identical particles—the method of quantized waves and that of the configuration space—can easily be traced in the case of the symmetrical functions. In fact, as we have just remarked, symmetrical functions are compatible with an individual state being occupied by any number of particles—this number being interpreted in the method of quantized waves as the quantum number of the corresponding harmonic oscillator. According to the usual theory of the quantization of such an oscillator, it is, however, impossible to limit this quantum number to the values 0 and 1, which is necessary for the quantized-wave-theory interpretation of the antisymmetric functions. To interpret these functions we must assume that the 'oscillators' (three-dimensional waves) representing such functions are quantized according to some peculiar rule which automatically excludes any values of their quantum numbers larger than 1. Such a rule has been worked out by Jordan and will be considered later (Part II). For the present we merely record the

existence of a fundamental distinction between the two types of quantized waves, corresponding to the symmetrical and antisymmetrical functions.

24. Relation between Matter and Light

We have already pointed out at the close of the preceding chapter a few distinctions between light and matter, which indicate that the analogy between them, which was the starting-point of the development of wave mechanics, is superficial and of very limited scope. We shall now examine more exactly the relation between light and matter from the point of view of the theory of quantized waves developed at the beginning of this chapter.

Light waves are usually described as electromagnetic waves, i.e. as waves of electrical and magnetic intensity $\mathbf{E}$, $\mathbf{H}$, and not as 'probability' waves, such as we have been using hitherto in connexion with the motion of a particle.

If we restrict ourselves to functions ($\mathbf{E}_k$, $\mathbf{H}_k$) or ψ_k associated with a definite stationary state (k), this distinction amounts to saying that the squares of the former define the *energy density*, while the square of the latter, or rather of its modulus, defines the *probability density*. This difference could be expressed by saying that the ($\mathbf{E}_k$, $\mathbf{H}_k$)-field represents 'energy waves', while the ψ_k-field represents 'probability waves'. It would obviously be possible to eliminate the distinction in question by dividing $\mathbf{E}_k$ and $\mathbf{H}_k$ by the square root of the energy quantum $\epsilon_k = h\nu_k$. Since in the case of light waves the electric and magnetic intensities are always numerically equal, we could drop the function H_k and define the energy density by $E_k^2/(4\pi)$ [instead of $(E_k^2+H_k^2)/(8\pi)$]. The wave function ψ_k, corresponding to $\mathbf{E}_k$, would then be given by the relation

$$\psi_k = \frac{\mathbf{E}_k}{\sqrt{(4\pi h\nu_k)}}. \tag{121}$$

This relation (implying that ψ_k must be considered as a vector) does not, however, take into account the fact that the ψ_k must be normalized, i.e. satisfy the condition $\int \psi_k \psi_k^* \, dV = 1$. Also it disregards the fact that ψ_k *must be a complex quantity* containing the time through the factor $e^{-i2\pi\nu_k t}$, so that $|\psi_k|^2$ should be independent of the time, whereas $\mathbf{E}_k$ must be a real quantity if its square is to be equal to the energy density, which is not a constant but oscillates with a frequency $2\nu_k$ about an average value $\overline{E_k^2}/(4\pi) = (E_k^0)^2/(8\pi)$, where E_k^0 is the amplitude of $\mathbf{E}_k$. The second difficulty can be formally removed by using, instead of $\mathbf{E}_k$, the corresponding complex quantity involving t through the

factor $e^{-i2\pi\nu_k t}$ and by replacing the actual energy density by its average value which, with the above condition, can be expressed in the form $\mathbf{E}_k \cdot \mathbf{E}_k^*/(8\pi)$. As for the first difficulty, we may avoid it by putting

$$\mathbf{E}_k = c_k \sqrt{(8\pi h\nu_k)}\psi_k, \tag{121a}$$

where c_k is a complex number whose absolute value is to be defined by the condition $|c_k|^2 = N_k$.

So far the distinctions between $\mathbf{E}$-waves and ψ-waves do not seem to be particularly important, at least from the purely formal point of view, although, as we are going to see presently, there are grave physical objections against considering E as a complex quantity similar to ψ.

The situation assumes a different aspect if we pass from waves corresponding to some particular state k to those associated with a number of different states, supposed to be realized simultaneously (for 'different' particles of course, since these states are alternative and *one* particle cannot simultaneously exist in all of them).

In this case the electric intensity $\mathbb{E}$ is represented by the sum of all the particular $\mathbf{E}_k$. Replacing the latter by unit intensities $\mathbf{e}_k$, corresponding to $N_k = 1$ and normalized according to the equation

$$\int \frac{|E_k|^2}{8\pi}\, dV = h\nu_k$$

(which is equivalent to $\int |\psi_k|^2\, dV = 1$), we have $\mathbf{E}_k = c_k\, \mathbf{e}_k$, and consequently

$$\mathbb{E} = \sum c_k\, \mathbf{e}_k. \tag{122}$$

This equation is similar to equation (105 b) for the wave function $\Psi = N\psi$—the $\mathbf{e}_k$ merely takes the place of the ψ_k. *There is, however, no relation whatever between the resulting* $\mathbb{E}$ *and the resulting* Ψ, for the relations between the particular $\mathbf{e}_k$ and the corresponding ψ_k,

$$\mathbf{e}_k = \sqrt{(8\pi h\nu_k)}\psi_k, \tag{122a}$$

involve coefficients which are different for the different particular states and of a type which excludes the possibility of establishing any general relation between $\mathbb{E}$ and ψ. It may be mentioned that if the coefficients in (122 a) were proportional to the frequencies instead of to the square roots of the frequencies, we could put $\mathbf{e}_k \sim d\psi_k/dt$ and $\mathbb{E} \sim d\Psi/dt$. As it is, we can use a symbolic relation of the same type involving the differential coefficient (with respect to the time) of the order $\frac{1}{2}$, i.e. we can put $\mathbf{e}_k = 2\sqrt{(hi)}\left(\frac{d}{dt}\right)^{\frac{1}{2}}\psi_k$, and accordingly

$$\mathbb{E} = 2\sqrt{(hi)}\left(\frac{d}{dt}\right)^{\frac{1}{2}}\Psi. \tag{122b}$$

It seems, however, doubtful whether this relation between $\mathbf{E}$ and Ψ has any physical significance.†

It follows from the preceding considerations that, in the case of light, there is no general connexion between the waves *considered as electromagnetic* waves and the particles or photons, in the sense that the wave function $\mathbf{E}(x, y, z, t)$ is not directly connected with the probability of locating a photon at the corresponding point—unless it reduces to a function $\mathbf{E}_k$ associated with a particular state.

The true connexion between the light waves and the probability waves of the electrons becomes particularly clear and spectacular in the general case of superposition of several harmonic electron waves with different frequencies, the product $\psi\psi^*$ containing 'interference terms' with frequencies $\nu_{mn} = \nu_m - \nu_n$. We have seen that these 'difference frequencies' have a very simple and deep physical meaning, since they represent the frequencies of the light that can be emitted by the electron in transitions from one state to another. This relation can be interpreted as the result of the proportionality between $e\psi\psi^*$, i.e. the probable density of the electrical charge e connected with the particle whose motion is described by the function ψ, and the (probable) electrical field E produced by this particle. We thus see that E corresponds not to the complex quantity ψ but to the real quantity $\psi\psi^*$. The fact that E must be a *real* quantity, in contradistinction to ψ, can be also interpreted as the mathematical expression of the fact that it represents a physically measurable quantity, while ψ does not, serving but for the calculation of the physically measurable $\psi\psi^*$. It is argued sometimes that for light-vibrations we can measure the average value of the energy only, and not the field strength as a function of the time. But on the other hand, for electromagnetic oscillations of moderate frequency we can certainly measure the field strength itself and follow its variation with the time (with the help of an oscillograph for example). It is true that the quantity directly measured is not the field strength itself but the force, that is, the product of E by the electrical charge

† For plane waves represented by the equation $E = E^0 \exp i2\pi(\boldsymbol{\kappa} \,.\, \mathbf{r} - \nu t)$ the relation (122 b) can be replaced by another symbolic relation $E = 2\sqrt{(-ihc)}\ \nabla^{\frac{1}{2}}\Psi$. Here $\nabla^{\frac{1}{2}}$ denotes a differentiation of the order $\frac{1}{2}$ with respect to the coordinates, equivalent to the multiplication of ψ by the square root of the magnitude of the vector $(i2\pi\boldsymbol{\kappa})$, whose scalar product with $\mathbf{r}$ appears in the exponent of the wave function. The equivalence of the two formulae follows from the relation $\nu = c\kappa$ where c is the velocity of light and $\kappa = 1/\lambda$ the wave number. The second formula can be generalized with the help of Fourier's theorem for waves of any type. It has actually been used recently by a few authors in a futile attempt to build up a 'quantum electrodynamics' which should treat photons in the same way as electrons are dealt with in the wave-mechanical theory.

of the particle upon which it is supposed to act. But this charge can be considered, for electrons for example, as a known quantity, and on the other hand, the force acting on it *cannot* be transformed into an expression quadratic in the resulting field strength (due both to the external field and the particle itself) and corresponding to the electromagnetic energy of the resulting field. It is therefore erroneous to think that the measurement of electric force can be reduced to the measurement of electromagnetic energy. On the same ground we might deny the possibility of measuring the electromagnetic energy, for from the point of view of the modern theory of gravitation it determines the gravitational field in a way similar to that in which the electric charge determines the electromagnetic field, the gravitational field strength being proportional to the mass, i.e. to the electromagnetic energy divided by the square of the velocity of light.

The similarity of the relation between the ψ-field and the E-field on the one hand, and of the relation between the latter and the gravitational field on the other, might be considered as a new form of the analogy between matter and light, replacing the superficial analogy which is based on the idea of wave-corpuscular parallelism. It does not imply the necessity of introducing the notion of photons as particles associated with the electromagnetic waves in the same way as electrons are associated with ψ-waves. Such particles are in fact no more necessary than 'gravitational quanta' or 'gravons' which we could associate with gravitational waves (or sound quanta or 'phonons' that could be associated with sound waves).

It should be mentioned that, with a real value of the electromagnetic field strength and an electromagnetic energy density proportional to its square, we must get, in the case of the superposition of harmonic electromagnetic oscillations with different frequencies ν_m and ν_n, terms both with the frequency $\nu_m - \nu_n$ and $\nu_m + \nu_n$ in the expression of the energy density, giving rise to gravitational waves (too weak of course to be actually observed) both with the difference frequency—as in the case of light waves produced by the interference of cathode waves—and with the sum frequency $\nu_m + \nu_n$ which from the corpuscular point of view should require an entirely new quantum interpretation.

The limited scope of the wave-mechanical theory of light, as constituted of particles associated through the conception of probability with waves, seems to be due to the fact that it overlooks the real *physical connexion* between light and matter—a connexion that can be

described as an analogy (in the sense of the wave-corpuscular parallelism) in a very superficial and incomplete way.

This should not astonish us if we remember that light is, after all, merely a particular case of an electromagnetic field of force describing the mutual action of the elementary particles of matter. It would be very strange indeed if this electromagnetic field displayed properties perfectly similar to those of matter itself (inasmuch as the latter is visualized as a system of particles). The lack of any actual similarity between them is very impressively demonstrated by comparing matter with another special case of the electromagnetic field—namely, the *electrostatic* field. The electrostatic field and the electromagnetic field of light waves can be considered as opposite extremes in the sense that the former is prevalent at small and the latter at large distances from material particles. The similarity between the propagation of the electromagnetic energy and the electromagnetic momentum associated with the light waves and the motion of ordinary matter (in the form of a continuous medium) can be established on the basis of a purely undulatory theory of light—a similarity which, though certainly very superficial and incomplete, has made it possible to develop a corpuscular theory of light.†

It thus seems to follow from the preceding considerations that this corpuscular—or wave-mechanical—theory of light is in itself wrong. Its main value is that it stimulated the development of the wave mechanics of matter. It may be remarked that it has been the lot of many wrong physical theories to pave the way for better ones.

† The energy flow is defined by the vector (Poynting's vector) $\frac{c}{4\pi}\mathbf{E}\times\mathbf{H}$. In the special case of the light waves, which are characterized by the numerical equality and mutual perpendicularity of $\mathbf{E}$ and $\mathbf{H}$, this vector reduces to the product of $\mathbf{c}$ (a vector in the direction of the propagation of the waves and numerically equal to the velocity of light c) by $\frac{E^2}{4\pi}=\frac{E^2+H^2}{8\pi}$, i.e. by the energy density, *just as if* the energy was transported by particles moving with the velocity $\mathbf{c}$. This analogy breaks down, however, when $\mathbf{E}$ and $\mathbf{H}$ are not numerically equal or not perpendicular to one another.

V

STATISTICAL MECHANICS

25. Statistical Equilibrium of a Gas at the Zero Point of the Temperature

We have hitherto confined ourselves to investigating the *possible* distributions of a system of identical particles, whose mutual action is negligible (or represented by an adequate external field of force), among the different individual states, and have found two types of possible distributions—exclusive distributions which correspond to antisymmetrical functions, and distributions which correspond to symmetrical functions and which in future will be referred to as '*inclusive*' distributions.

We shall now examine those distributions—exclusive or inclusive—which *actually* occur when the (nearly) ideal gas, formed by the particles under consideration, is in a state of 'statistical' or 'thermal' equilibrium.

The simplest of these actual distributions is that corresponding to *the absolute zero of temperature*. It is determined by the condition that the total energy of the system, i.e. the sum of the energies of the separate particles (since their mutual action is either neglected or accounted for by a suitable external field), must have the least possible value. In the classical statistical theory, which ignores the identity principle and regards all distributions as possible, this condition is satisfied by assuming that, at the absolute temperature $T = 0$, all the particles are in the same individual state, namely, the state of lowest energy, which is the 'normal' state for each of them taken separately.

The same result holds for the inclusive distribution which only begins to differ from the classical at temperatures greater than zero. The exclusive distribution, however, is quite different. For instance, for an electron gas of N electrons at $T = 0$ the N electrons are associated one by one with each of the N lowest states.

The normal distribution of the electrons in an atom (other, of course, than that of hydrogen) is an example of this type of distribution. If each pair of states differing only by the 'spin quantum number' (i.e. by the orientation of the electrons) is regarded as a double state, then the normal distribution is obtained by associating *two* electrons with each of the $\frac{1}{2}N$ lowest double states. This means that the energy of the system, or, what amounts to the same thing, the *average* energy of a single particle, at zero temperature is larger than that required by

classical statistics. In particular, for an ideal gas the classical energy would be equal (practically) to zero, all particles being supposed to be at rest. But in the exclusive distribution of an electron gas at absolute zero, only two electrons can remain at rest, or more exactly in the state of lowest energy; the others will have to accommodate themselves by pairs among the states of higher energy. This picture suggests the analogy to a big apartment house, whose inhabitants tend to crowd into the lower floors, but are restricted in this by a law, providing that each double room should be occupied by one married couple only. The opposite sexes correspond to the opposite direction of the 'spins' of the electrons. To make the analogy more complete we should imagine the whole building to consist of such double-room apartments (there can be many of them on the same floor corresponding to a given energy-level) and the number of floors to be infinite. This 'skyscraper' can be imagined to be erected in the phase-space of classical statistical mechanics, and all the 'double-room apartments'—or elementary cells as they are usually called—to have exactly the same volume h^3. These 'cells' correspond to stationary states specified by the three ordinary quantum numbers n_x, n_y, n_z, or n_1, n_2, n_3, the value of the spin number remaining unspecified.

The above analogy serves to bring out more clearly the remarkable fact that the 'different' electrons, despite their supposed lack of dynamical interaction, are statistically interconnected with one another so as to conform to Pauli's 'exclusion law'. They behave as if they were aware of one another's presence and, in spite of their tendency to crowd into the more comfortable lower rooms of the big apartment house built up in the phase-space, they never intrude in an apartment already occupied by some happier electron couple.

It may seem at first that such considerate behaviour by the electrons is physically impossible and that it can be interpreted only by ascribing to them some 'extra-physical' properties, equivalent to the 'tendency' of our metaphorical population to concentrate at the bottom of the 'state-building' on the one hand and their respect for the restrictions imposed upon them by the exclusion law on the other. As a matter of fact such extra-physical properties are not necessary. The tendency of any system of particles that is not absolutely isolated from the external world towards the state of lowest energy at the temperature of the absolute zero is a consequence of the definition of the absolute zero of temperature in connexion with the general law of statistical or thermal equilibrium. As to the exclusion law, although its origin is not clear,

yet if in operation at a certain instant it continues to apply for all time in spite of any perturbing influences. As we have seen in the preceding chapter, this can be deduced from the laws of wave mechanics in conjunction with the identity principle, which itself is a corollary from these laws.

26. General Law of Statistical Equilibrium

The idea of temperature cannot be introduced in a general way before but only after the character of the equilibrium distribution has been determined, the temperature then entering as a parameter in the distribution law.

If the system of particles under consideration is treated as one whole, then this law can be stated in the following general form: The probability that the system will be found in a certain state α is determined by the corresponding value of the total energy E_α according to the formula

$$P_\alpha = Ae^{-E_\alpha/\theta}, \tag{123}$$

where A and θ are constants, independent of α.

Formula (123) was established in classical statistical mechanics by Boltzmann and bears his name. The constant θ is a measure of the absolute temperature T. The latter is usually defined by the relation

$$\theta = kT, \tag{123 a}$$

where $k = 1{\cdot}37 \times 10^{-16}$ erg/deg. is called Boltzmann's constant. The constant A is determined as follows: The system must be in one of its possible states. Therefore the sum of the probabilities taken over all possible states must be equal to 1. Applying the classical addition theorem, we get $A \sum_\alpha e^{-E_\alpha/\theta} = 1$ or

$$A = \frac{1}{\sum_\alpha e^{-E_\alpha/\theta}}. \tag{123 b}$$

The preceding results can be shown in a very general way to remain valid in the new statistical theory based on wave mechanics.

In order to establish them we must first clearly define the fundamental idea of *statistical or thermal equilibrium* to which these results refer.

The idea of equilibrium is a relative one. It implies the existence of several systems or parts of a system which are in equilibrium *with respect to one another* (just as the idea of potential energy implies the existence of systems or particles which act on one another). The

relative equilibrium of two systems a and b can, however, be defined only when their mutual action is *comparatively weak*, so that a and b behave approximately as if they were completely isolated from one another. This condition can be stated in a more exact form as follows: the mutual (potential) energy E_{ab} of a and b must be very small compared with the internal energy E_{aa} of a (including the mutual potential energy of the particles constituting a, if any) and compared with the internal energy E_{bb} of b. Nevertheless, E_{ab} must not be zero, for then, i.e. in the complete absence of any interaction between a and b, there can be no question of equilibrium between them.

Thus, in order to find the equilibrium condition for the system a under consideration, we must *either* subdivide it into smaller parts—which may be the separate particles—*or* consider it as a part of a larger system c, whose other part will be denoted by b.

The result stated above (Boltzmann's distribution law) can be obtained by the second method, which we shall examine first because of its greater generality.

Because of the supposed weakness of its mutual action with b, the states of the system a will be the same as if it were completely isolated from external influences. The same of course is true of b. The effect of the interaction between a and b will then be to cause transitions from the initial states α' and β' (say) of a and b into other states α'', β'' or α''', β''', and so on, which satisfy the condition that the sum of the energies of a and b remains constant:

$$E_{\alpha'}+E_{\beta'} = E_{\alpha''}+E_{\beta''}. \tag{124}$$

The energy E_{ab} will not enter into this balance equation—not merely because of its relative smallness but because it is already taken into account as the cause of transitions in the system (a, b).

Thanks to these transitions the state of a must be represented wave-mechanically as the superposition of a number of alternative stationary states α', α'',.... If the latter are associated with harmonic functions $\psi_{\alpha'}$, $\psi_{\alpha''}$,... normalized in the usual way ($\int |\psi_\alpha|^2\, dV_a = 1$), then the function representing the behaviour of a will be

$$\psi_a = c_{\alpha'}\psi_{\alpha'}+c_{\alpha''}\psi_{\alpha''}+\ldots,$$

where $|c_\alpha|^2$ is the relative probability of finding a in a state α (cf. Ch. III, § 17). Similar results hold of course for b. The behaviour of the combined system (a, b) will thus be described by the product

$$\psi_{ab} = \psi_a\psi_b = c_{\alpha'\beta'}\psi_{\alpha'}\psi_{\beta'}+c_{\alpha''\beta'}\psi_{\alpha''}\psi_{\beta'}+\ldots,$$

where $|c_{\alpha'\beta'}|^2 = |c_{\alpha'}|^2|c_{\beta'}|^2$ is the probability of simultaneously finding a in the state α' and b in the state β'.†

Instead of speaking of probabilities it is more convenient to think of a large number of 'copies' of the system (a, b). We may then interpret $|c_\alpha|^2|c_\beta|^2$ as the relative number of these copies for which a is in the stationary state α and b is in the stationary state β. With this convention we can talk of the (relative) number of transitions between various (compounded) states (α, β) of the system (a, b) in a given time, and can define statistical equilibrium as that condition *in which the number of transitions between any two states* (α', β') *and* (α'', β'') *is the same for the two opposite directions in which these transitions can take place.*

The number of transitions from the state (α', β') to the state (α'', β'') must obviously be proportional to $|c_{\alpha'}|^2|c_{\beta'}|^2$ which represents the probability of simultaneously finding a in α' and b in β', or, in other words, which represents the relative number of systems (a, b) in the state (α', β'). In addition the number of transitions is proportional to the transition probability $f_{(\alpha'\beta')\to(\alpha''\beta'')}$ which, as we know, is measured by the square of the modulus of the perturbation energy E_{ab} with respect to the functions $\psi_{\alpha'}\psi_{\beta'}$ and $\psi_{\alpha''}\psi_{\beta''}$. The number of transitions taking place in the same time in the opposite direction will be proportional to the product $|c_{\alpha''}|^2|c_{\beta''}|^2 f_{(\alpha''\beta'')\to(\alpha'\beta')}$, the coefficient of proportionality being the same in both cases.

It has been shown in Ch. III, § 18, that the transition probabilities in the two directions $f_{(\alpha'\beta')\to(\alpha''\beta'')}$ and $f_{(\alpha''\beta'')\to(\alpha'\beta')}$ are also equal; whence it follows that the condition for equilibrium can be stated in the form

$$|c_{\alpha'}|^2|c_{\beta'}|^2 = |c_{\alpha''}|^2|c_{\beta''}|^2. \tag{124 a}$$

Comparing this equation with (124) we see that it can only be satisfied when the quantities $|c_\alpha|^2$ and $|c_\beta|^2$ are given by

$$|c_\alpha|^2 = Ae^{-E_\alpha/\theta}, \qquad |c_\beta|^2 = Be^{-E_\beta/\theta}$$

with the same value of the parameter θ. The latter condition expresses the equality of the temperatures of a and b. If we write P_α for $|c_\alpha|^2$ we obtain Boltzmann's equation (123).

It follows from this equation that for $\theta = 0$, i.e. at the absolute zero of temperature, one state only is possible, namely, the state of *lowest* energy. In fact the relative probability of two states α' and α'' is given

† This multiplicative representation of ψ_a, if valid at some initial instant, will be approximately maintained so long as the perturbation energy E_{ab} is small. In the general case the coefficient $c_{\alpha\beta}$ is not equal to the product of c_α and c_β.

by the ratio $P_{\alpha''}/P_{\alpha'} = e^{-(E_{\alpha''}-E_{\alpha'})/\theta}$ which reduces to zero if $\theta = 0$ and $E_{\alpha''} > E_{\alpha'}$.

It should be noticed here that the 'zero point' condition, in the sense of the prevalence of the state of lowest energy, is practically reached at temperatures for which the value of θ is small compared with the 'excitation energy' of the system under consideration, i.e. the energy difference between the normal state and the next lowest state on the energy scale. In many cases (for instance, for atoms) this excitation energy is so large that the zero-point condition is practically maintained up to temperatures of tens and even hundreds of thousands of degrees.

27. Application of the General Law of Statistical Equilibrium to an Ideal Gas

We shall now use the preceding results to determine the equilibrium distribution in an ideal gas† consisting of N identical particles. A distribution will be specified by giving the number of particles N_1, N_2,..., N_k,... which are to be found in the different individual states 1, 2,..., k,.... Each of the latter numbers represents the group of quantum numbers specifying the respective state. For instance, k stands for the quantum numbers $n_x^{(k)}$, $n_y^{(k)}$, $n_z^{(k)}$, $s^{(k)}$. Our task is to determine the *average values* of the partition numbers N_i corresponding to the expression (123) for the probabilities of the various (aggregate) states of the whole system. These average values of the N_k will be denoted by $\overline{N}_k$. They do not, strictly speaking, correspond to a definite permanent distribution, for actually there is no such thing as a permanent distribution. In reality the number of particles associated with each state, instead of being constant, rapidly and irregularly oscillates ('fluctuates') about the average value $\overline{N}_k$. These oscillations or *fluctuations* can be estimated by computing the average values of the squares of the differences $N_k - \overline{N}_k$. The equilibrium distribution is therefore defined as the *average distribution* of the particles among the individual states, corresponding to the probability law (123) for the resulting aggregate states.

The numbers α labelling these states represent a definite set of values (including zero) for the partition numbers $N_1^{(\alpha)}$, $N_2^{(\alpha)}$,..., $N_k^{(\alpha)}$,.... Since the interaction of the particles is neglected, the energy E_α of the combined state α is therefore given by

$$E_\alpha = \sum_k N_k^{(\alpha)} W_k, \tag{125}$$

† By an ideal gas we mean a system of particles which are supposed not to act on one another; they can, however, be acted upon by external forces, including a field of force representing approximately their mutual action.

where W_k is the energy of a single particle in the corresponding individual state, and the summation is extended over *all* such states.

The average value of any function F of α (i.e. of the numbers $N_k^{(\alpha)}$) is given by

$$F = \sum_\alpha F_\alpha P_\alpha,$$

or, according to (123) and (123 b),

$$\bar{F} = \frac{\sum F_\alpha e^{-E_\alpha/\theta}}{\sum e^{-E_\alpha/\theta}}. \tag{125 a}$$

In carrying out the summation three cases must be distinguished:

(1) *Classical* ('*individualistic*') *statistics*. The identity principle is here not recognized and the 'different' particles therefore retain their individuality. A distribution specified by the partition numbers $N_1^{(\alpha)}$, $N_2^{(\alpha)}$,... is therefore considered not as a single state α of the whole system but as a set of equivalent states differing from one another with respect to the individuality of the particles assigned to the various individual states. The number of these equivalent states is given by

$$g_\alpha = \frac{N!}{N_1!\, N_2! \ldots}, \tag{126}$$

which is equal to the number of different arrangements of N *different* things among a number of boxes in such a way that there are N_k things of any kind in the kth box.†

Thus in equation (125 a) each term under the sign of summation must be repeated g_α times, or, in other words, $e^{-E_\alpha/\theta}$ must be replaced by the product $g_\alpha e^{-E_\alpha/\theta}$.

(2) *The new 'quantum' or 'collectivistic' statistics of the symmetrical type* called the *Bose-Einstein* statistics (Bose having introduced it in studying photons and Einstein in studying particles of ordinary matter). In this case each state α, corresponding to *any* distribution, is to be counted *once*. The difference between the symmetric or 'inclusive' statistics and classical statistics is vividly shown if, as before, we use a factor g_α to represent the number of times each state α is to be repeated. Instead of formula (126) we now have

$$g_\alpha = 1. \tag{126 a}$$

(3) *The new 'collectivistic' statistics of the antisymmetrical type* is called the Fermi-Dirac statistics or, more correctly, the statistics of *Pauli, Fermi, and Dirac*. In this case the only possible states α are those for which the partition-numbers $N_k^{(\alpha)}$ are either 1 or 0. This

† It is understood that $N_k! = 1$ if $N_k = 0$.

restriction is imposed by the exclusion principle and is equivalent to multiplying each term $e^{-E_\alpha/\theta}$ in the unrestricted summation by a factor

$$g_\alpha = \begin{cases} 1 & \text{if all the } N_k^{(\alpha)} = 0, 1 \\ 0 & \text{if any one of the } N_k^{(\alpha)} > 1. \end{cases} \tag{126 b}$$

The factors g_α defined by the above formulae are usually known as the '*weight factors*' of the respective aggregate states. Introducing these factors, we can replace (125 a) by the expression

$$\overline{F} = \frac{\sum F_\alpha g_\alpha e^{-E_\alpha/\theta}}{\sum g_\alpha e^{-E_\alpha/\theta}}, \tag{127}$$

where *each* state α is supposed to occur in the summation once and *once* only.

To calculate the required average distribution we should replace F_α in equation (127) by $N_k^{(\alpha)}$. It is simpler, however, first to calculate the average value of the total energy E. Putting $F_\alpha = E_\alpha$ and noticing that

$$E_\alpha e^{-E_\alpha/\theta} = -\frac{\partial}{\partial \mu}(e^{-\mu E_\alpha}),$$

where $\mu = 1/\theta$, we get

$$\overline{E} = -\frac{\frac{\partial}{\partial \mu}\sum g_\alpha e^{-\mu E_\alpha}}{\sum g_\alpha e^{-\mu E_\alpha}},$$

or

$$\overline{E} = -\frac{\partial}{\partial \mu}\log Z, \tag{128}$$

with

$$Z = \sum g_\alpha e^{-\mu E_\alpha}. \tag{128 a}$$

The quantity Z will be referred to in future as the *partition function* of the system in question (the 'Zustandssumme' of the German authors). It is equal to the reciprocal of the coefficient A in (123), or rather in the 'weighted' expression $P_\alpha = A g_\alpha e^{-\mu E_\alpha}$, μ being the reciprocal of θ or kT.

The evaluation of (128 a) is very easy in the classical ('individualistic') statistics; it simply reduces to the Nth power of

$$z = \sum_k e^{-\mu W_k} \tag{128 b}$$

summed over all the states of a single particle. In fact by the well-known multinomial theorem we have:

$$\left(\sum_k x_k\right)^N = \sum \frac{N!}{N_1!\,N_2!\ldots} x_1^{N_1} x_2^{N_2}\ldots$$

which, with $x_k = e^{-\mu W_k}$, reduces to (128 a) in consequence of (125). We thus get

$$Z = z^N,$$

whence, according to (128),

$$\bar{E} = -N\frac{\partial}{\partial\mu}\log z = N\frac{\sum W_k e^{-\mu W_k}}{\sum e^{-\mu W_k}}.$$

On the other hand it follows directly from (125) that

$$\bar{E} = \sum_k \bar{N}_k W_k.$$

Comparing this with the preceding expression, we obtain the result

$$\bar{N}_k = \frac{Ne^{-\mu W_k}}{\sum_k e^{-\mu W_k}} = \frac{N}{z}e^{-W_k/\theta}, \tag{129}$$

which is quite similar to the original equation (123) for the whole system. The distribution defined by (129) is a complete statement of Boltzmann's (or the Maxwell-Boltzmann) law.

The states of classical statistics are not quantized and therefore the 'individual states' labelled k have there no definite meaning. We may give them a meaning by noticing that a quantized state corresponds to a definite volume of the classical 'phase-space', namely

$$\int \ldots \int dxdydz\, dg_x dg_y dg_z = h^3. \tag{130}$$

In going over to classical statistics we can still retain the idea of quantization *by considering h, not as a given universal constant, but as an infinitesimal quantity*, so that the successive quantized states lie infinitely near to one another. Quantities such as the energy W, which take definite quantized values W_k for the different quantized states, *can now be treated as continuous functions of the variables x, y, z, g_x, g_y, g_z*, just as in classical mechanics where there is no quantization at all. The expression (128b) for z can therefore be replaced by the integral

$$z = \frac{1}{h^3}\int e^{-\mu W}\, d\gamma, \tag{131}$$

where $d\gamma$ denotes the element of volume of the phase-space of an individual particle ($d\gamma = dxdydz\, dg_x dg_y dg_z$) and the integration is extended over all that region of the *phase-space in which the particle is allowed to move.*

Expression (129) now becomes

$$dN = N\frac{e^{-W/\theta}\, d\gamma}{\int e^{-W/\theta}\, d\gamma}, \tag{131a}$$

where dN denotes the number of particles whose states, in the classical sense of the term, are contained within the element of volume $d\gamma$ of the

phase-space. We see, by the way, that the constant h^3 appearing in (131) no longer appears in (131a). We may therefore drop it from (131) also, since it has no effect on the distribution law. The quantity z is then defined merely as the integral $\int e^{-W/\theta}\, d\gamma$ ('partition integral' for a single particle).

The transition to the classical theory may be made somewhat earlier when calculating the quantity Z referring not to a single particle, but to the whole system. Denoting the phase-spaces of the individual particles by $\gamma_1, \gamma_2, \ldots, \gamma_N$, we can specify the state of the whole system (in the classical sense) by the volume element $d\Gamma = d\gamma_1 d\gamma_2 \ldots d\gamma_N$ of the resulting phase-space, a quantized state of the system corresponding obviously to a volume

$$\int d\Gamma = h^{3N}$$

of this space. By letting h tend to zero we can, as before, actually remove the quantization. If in addition the particles are treated as different, so that the weight-factor g_α is defined by (126), we can put simply

$$Z = \frac{1}{h^{3N}} \int e^{-\mu E}\, d\Gamma, \tag{132}$$

because the integration over Γ automatically secures the repetition of each distribution, specified by the partition numbers $N_1, N_2, \ldots$, just $\frac{N!}{N_1!\, N_2! \ldots}$ times. Actually, because h is supposed to be infinitely small, the elements of volume $d\gamma_1, d\gamma_2, \ldots, d\gamma_N$ are also infinitely small, and therefore the chance of some of these elements of volume coinciding, when the individual phase-spaces are imagined to be superposed, is so small that it can be neglected. The numbers N_k must therefore be either 1 or 0 and the weight-factor g reduces to the constant value $N!$.

Turning now to the determination of the numbers $\overline{N}_k$ in the new quantum statistics we must preserve the original definition of Z as a sum and use the expressions (126a) or (126b) to determine the weight-factor g. Because of the indistinguishability of the particles, the sum Z no longer reduces, as it does in classical statistics, to the product of factors depending on the separate particles, and the exact calculation of Z proves to be very difficult.

This difficulty can be overcome, as shown by Pauli, by using the following artifice, originally used in a similar connexion by Gibbs. The given system with a fixed number of particles N is replaced by an infinite number of similar systems differing from one another by the values of N in such a way that the relative number of systems with a

given N *and a definite energy* E is proportional to $e^{-\lambda N}$. In other words, the probability of finding a system consisting of N particles and having an energy E is defined by

$$P = Ae^{-\lambda N - \mu E}, \tag{133}$$

where A is a constant independent both of N and E, and λ is a constant defining the average number of particles $\overline{N}$ in a way similar to that in which μ, the reciprocal of θ, defines the average energy $\overline{E}$.

Formula (133) can be considered as a generalization of Boltzmann's law (123) and can be derived in a somewhat similar way by allowing not only an exchange of energy, but also *an exchange of particles* between the two parts a and b in which a given closed system c is supposed to be divided. Since the total number of particles, as well as the total energy, is fixed, we now have two conservation equations, i.e.

$$N_{\alpha'} + N_{\beta'} = N_{\alpha''} + N_{\beta''}, \qquad E_{\alpha'} + E_{\beta'} = E_{\alpha''} + E_{\beta''}.$$

The probability equation

$$P_{\alpha'} P_{\beta'} = P_{\alpha''} P_{\beta''},$$

which expresses the condition for equilibrium between the transitions $(\alpha'\beta') \rightarrow (\alpha''\beta'')$ and $(\alpha''\beta'') \rightarrow (\alpha'\beta')$, is the same as before. It can easily be shown that this probability equation is consistent with the conservation equations only when $P_\alpha = Ae^{-\lambda N_\alpha - \mu E_\alpha}$ and $P_\beta = Be^{-\lambda N_\beta - \mu E_\beta}$, the constants λ and μ being the same for a and b. Dropping the index α we get Pauli's formula (133).

This derivation of (133) is, however, based on the tacit assumption that the number of (identical) particles constituting a *given* system is not an essential constant for that system but a variable of the same sort as the energy. This assumption can be justified if the system is described by the method of quantized waves (§ 20), the number N_α being defined as the sum of the quantum numbers N_1, N_2,... which refer to the oscillators representing the different quantized states of a single particle.

The average values of N and E (for a) are given, according to (133), by the equations

$$\overline{N} = -\frac{\partial}{\partial \lambda} \log Z, \qquad \overline{E} = -\frac{\partial}{\partial \mu} \log Z, \tag{133 a}$$

with

$$Z = \sum_N \sum_E e^{-\lambda N - \mu E}. \tag{133 b}$$

(We have dropped the weight-factor g, and in doing so we assume that only allowed distributions are to be taken into account.)

Let us first consider 'inclusive' distributions of the symmetrical type, i.e. distributions corresponding to $g = 1$ for any values of the partition numbers $N_1, N_2, \ldots$.

Putting $N = \sum_k N_k$ and $E = \sum_k N_k W_k$ (the index k referring, as before, to the different quantized states of a single particle), we can write Z in the form

$$Z = \sum_{N_1, N_2, \ldots} e^{-\sum_k (\lambda + \mu W_k) N_k},$$

where $\sum_{N_1, N_2, \ldots}$ denotes a summation over all the values of each of the numbers N_k from 0 to ∞ regardless of the value of the total number of particles $N = \sum_k N_k$. It is just this independence of the different N_k (achieved by dropping the restriction $\sum_k N_k = \text{const.}$) which makes the calculation of Z easy. In fact Z now reduces to the product of the sums

$$\sum_{N_k=0}^{\infty} e^{-(\lambda+\mu W_k)N_k} = \frac{1}{1 - e^{-(\lambda+\mu W_k)}}$$

referring to each of the individual states taken separately.† We thus get

$$\log Z = -\sum_k \log[1 - e^{-(\lambda+\mu W_k)}], \tag{134}$$

and hence, by (133 a),

$$\bar{N} = \sum_k \frac{1}{e^{\lambda+\mu W_k} - 1}, \tag{134 a}$$

$$E = \sum_k \frac{W_k}{e^{\lambda+\mu W_k} - 1}.$$

These expressions show that the average number of particles in the different (individual) states is given by the following 'equilibrium distribution law':

$$\bar{N}_k = \frac{1}{e^{\lambda+\mu W_k} - 1}. \tag{134 b}$$

Similarly, in the case of the exclusive distributions of the antisymmetrical type (corresponding to $g = 1$ for $N_k = 0, 1$ and to $g = 0$ for $N_k > 1$), we get

$$Z = \prod_k \sum_{N_k=0}^{1} e^{-(\lambda+\mu W_k)N_k} = \prod_k [1 + e^{-(\lambda+\mu W_k)}]$$

or

$$\log Z = \sum_k \log[1 + e^{-(\lambda+\mu W_k)}], \tag{135}$$

† To see this more clearly let us put for brevity $e^{-(\lambda+\mu W_k)} = x_k$.

Then $$Z = \sum_{N_1, N_2, \ldots} x_1^{N_1} x_2^{N_2} \ldots = \sum_{N_1} x_1^{N_1} \sum_{N_2} x_2^{N_2} \ldots = \prod_k \sum_{N_k=0}^{\infty} x_k^{N_k} = \prod_k \frac{1}{1 - x_k}.$$

whence $$\bar{N} = \sum_k \bar{N}_k, \qquad \bar{E} = \sum_k \bar{N}_k W_k, \tag{135a}$$

with $$\bar{N}_k = \frac{1}{e^{\lambda+\mu W_k}+1}. \tag{135b}$$

It may be worth while to check the above method by applying it to classical statistics $\left(\text{corresponding to } g = \frac{N!}{N_1!\,N_2!\ldots}\right)$ and comparing the results with those obtained before with a fixed value of N. Replacing the expression (133b) for Z by

$$Z = \sum \frac{N!}{N_1!\,N_2!\ldots} e^{-\lambda N-\mu E} = \sum_{N_1,N_2,\ldots} \frac{N!}{N_1!\,N_2!\ldots} x_1^{N_1} x_2^{N_2}\ldots,$$

with $x_k = e^{-\lambda-\mu W_k}$, we get

$$Z = \sum_N (x_1+x_2+\ldots)^N = \frac{1}{1-z},$$

making the natural and unavoidable assumption that the sum

$$z = \sum_k x_k = e^{-\lambda} \sum_k e^{-\mu W_k}$$

is smaller than 1 (otherwise z would have no definite value). Hence it follows from the equations (133a) that

$$\bar{N} = \frac{z}{1-z},$$

and

$$\bar{E} = \sum_k \frac{W_k x_k}{1-z},$$

and consequently $$\bar{N}_k = \frac{x_k}{1-z} = \text{const. } e^{-\mu W_k},$$

which is Boltzmann's distribution law. The last expression, as well as the corresponding expressions (134b) and (135b) for the average values of the partition numbers in the quantum statistics, can of course be obtained directly, i.e. without the help of the expression for the average value of the energy E.

We shall now consider the more general problem—namely, the determination of the average value of any function $F(N_k)$ of the partition number referring to a definite individual state k. This is done most simply with the help of the expression (133) for the probability of a distribution involving an arbitrary number of particles, the average value of F being defined by the formula

$$\bar{F} = \frac{\sum F e^{-\lambda N-\mu E}}{\sum e^{-\lambda N-\mu E}} \tag{136}$$

instead of (125a) or (127).

Inserting $N = \sum_k N_k$ and $E = \sum_k N_k W_k$, and using the fact that the numbers N_1, N_2,... can be dealt with as independent variables, we get

$$\overline{F(N_k)} = \frac{\sum_{N_k} F(N_k) e^{-\rho_k N_k}}{\sum_{N_k} e^{-\rho_k N_k}} \qquad (\rho_k = \lambda + \mu W_k); \tag{136 a}$$

all the factors containing partition numbers which refer to other individual states cancel out.

If $F(N_k) = N_k$, this gives

$$\overline{N_k} = -\frac{\partial}{\partial \rho_k} \log\Big(\sum_{N_k} e^{-\rho_k N_k}\Big); \tag{136 b}$$

since for classical statistics as well as for the statistics of Bose-Einstein N_k runs from 0 to ∞, we have

$$\sum_{N_k} e^{-\rho_k N_k} = \frac{1}{1 - e^{-\rho_k}},$$

and therefore

$$\overline{N_k} = \frac{\partial}{\partial \rho_k} \log(1 - e^{-\rho_k}) = \frac{1}{e^{\rho_k} - 1},$$

in agreement with (134 b). The classical case corresponds to the limit of large λ values.

In the Pauli-Fermi statistics we have, on the other hand,

$$\sum_{N_k} e^{-\rho_k N_k} = 1 + e^{-\rho_k},$$

and consequently, according to (136 b),

$$\overline{N_k} = \frac{1}{e^{\rho_k} + 1}$$

in agreement with (135 b).

Interesting results are obtained further by putting

$$F(N_k) = (N_k - \overline{N_k})^2 = (\Delta N_k)^2.$$

The average value of this expression determines the mean square deviation from the average (or most probable) distribution of the particles among the different individual states. Since

$$\overline{(N_k - \overline{N_k})^2} = \overline{N_k^2} - 2\overline{N_k}\,\overline{N_k} + (\overline{N_k})^2;$$

our task is reduced to the calculation of $\overline{N_k^2}$. This, according to (126 a), can be written in the form

$$\overline{N_k^2} = \frac{\dfrac{\partial^2}{\partial \rho_k^2} Z_k}{Z_k},$$

where Z_k denotes the sum $\sum_{N_k} e^{-\rho_k N_k}$. We thus get in the case of the

Bose-Einstein statistics

$$\overline{(\Delta N_k)^2} = \overline{N_k}(1+\overline{N_k}), \tag{137}$$

and in that of Pauli and Fermi

$$\overline{(\Delta N_k)^2} = \overline{N_k}(1-\overline{N_k}), \tag{137 a}$$

whereas the classical statistics is characterized by the relation

$$\overline{(\Delta N_k)^2} = \overline{N_k}, \tag{137 b}$$

to which (137) and (137 a) both reduce in the limiting case of small values of N_k.

The additional term $\pm(\overline{N_k})^2$ appearing in (137) and (137 a) corresponds to a statistical inter-dependence of the particles, which is a characteristic feature of the quantum statistics.†

It can be interpreted also, from the point of view of the wave mechanics as the result of the interference of the quantized waves, representing the particles under consideration, on the assumption that the waves corresponding to different numbers of particles are *incoherent* (i.e. devoid of definite phase relations) with regard to each other. This interpretation was first pointed out by Einstein, in connexion with the energy fluctuations in the radiation field, as described by Planck's formula.

28. Alternative Derivation of the Equilibrium Distribution of a Gas

In the foregoing section the equilibrium distribution was determined by considering the given system as part of a larger system. We shall now derive the same results by the alternative method, mentioned above, of dividing the system into parts and considering the equilibrium between these parts. In this subdivision of the system into parts we can and shall proceed to the ultimate parts, i.e. the particles which constitute it. This is possible for a gas because the mutual energy of

† The fact that the absence of such inter-dependence necessarily leads to the equation (137 b) can be shown by the following elementary derivation of the latter. Let the probability of a particle being found in the kth state be denoted by p_k. Then the probability of finding N_k *unspecified* particles in this state will be equal to

$$\frac{N!}{N_k!(N-N_k)!} p_k^{N_k}(1-p_k)^{N-N_k} = P_k.$$

The independence of the particles of each other is expressed here by the application of the multiplication law of probabilities ($1-p_k$ being the probability of finding a given particle in any state other than k). Using this expression for P_k we easily get

$$\overline{N_k} = \sum_{N_k} N_k P_k = N p_k \quad \text{and} \quad \overline{(\Delta N_k)^2} = \sum_{N_k} (\Delta N_k)^2 P_k = \overline{N_k}(1-\overline{N_k}/N)$$

which is practically equal to N_k so long as $N_k \ll N$.

the particles† may be assumed to be very small compared with the sum of their individual energies. The mutual energy may therefore be regarded as causing transitions of the particles between the various individual states. Hence in defining the states the mutual energy is to be neglected, just as in the treatment of the system (a, b). The equilibrium condition is then expressed, as before, by equating the number of transitions of a given type to the number of reverse transitions in the same time (principle of detailed balancing).

Instead of specifying the state of a particle exactly, it is more convenient to combine a number of neighbouring states into groups or *sets*. Let the number of stationary states in the ith set be g_i, and the number of particles belonging to this set be Q_i. The average number of particles associated with one of the states forming the ith set will therefore be equal to

$$N_i = \frac{Q_i}{g_i}.$$

Let us pick out a certain particle belonging to the i'th set and another belonging to the k'th set of states. Their interaction will cause them to effect transitions from their initial states into some other states (i'', k''). The probability $f_{(i'k')\rightarrow(i''k'')}$ of such a transition per unit time must be equal to the probability $f_{(i''k'')\rightarrow(i'k')}$ of the reverse transition. It is different from zero only when the total energy of the two particles (without their mutual energy) is the same for the initial and the final states. Since states belonging to the same set are supposed to have approximately the same energy, the energy of a state in the set i can be denoted by W_i. The energy condition just mentioned can therefore be written in the form

$$W_{i'} + W_{k'} = W_{i''} + W_{k''}. \tag{138}$$

According to the usual statistical conceptions, the number of transitions of the type $(i'k') \rightarrow (i''k'')$ in our 'nearly ideal' gas must be equal to

$$f_{(i'k')\rightarrow(i''k'')} Q_{i'} Q_{k'} g_{i''} g_{k''};$$

for $f_{(i'k')\rightarrow(i''k'')}$ is the probability of transition from a pair of *sharply* defined states in the sets i' and k' to another pair of sharply defined states in the sets i'' and k'', and to get all the transitions to this pair of sets we must multiply f by the numbers of states in each of them, and by the number of pairs of systems in the initial states. Equating the number of direct transitions to the number of reverse transitions defined in the same way, and cancelling the coefficients f because of

† Or that part of it not accounted for by an equivalent (self-consistent) external field.

their equality, we find for equilibrium

$$Q_{i'}\,Q_{k'}\,g_{i''}\,g_{k''} = Q_{i''}\,Q_{k''}\,g_{i'}\,g_{k'}$$

or

$$\frac{Q_{i'}}{g_{i'}}\cdot\frac{Q_{k'}}{g_{k'}} = \frac{Q_{i''}}{g_{i''}}\cdot\frac{Q_{k''}}{g_{k''}}, \tag{138 a}$$

which, combined with the energy condition (138), leads to the classical (Boltzmann) distribution law

$$\frac{Q_i}{g_i} = N_i = Ae^{-\mu W_i}.$$

To obtain the distribution laws corresponding to the wave-mechanical theory we must modify the preceding argument—the modification being obvious for the exclusive distribution. To begin with, the numbers Q_i must not exceed the corresponding 'state numbers' g_i, since the number of particles in each state must not be greater than 1. To maintain this restriction in spite of transitions, we must only allow transitions $(i'k') \to (i''k'')$ to states of the sets i'' and k'' which are unoccupied before the transition takes place. The number of such vacant states in the two sets are equal to $g_{i''}-Q_{i''}$ and $g_{k''}-Q_{k''}$ respectively. The total number of transitions $(i'k') \to (i''k'')$ will thus be proportional not to $Q_{i'}\,Q_{k'}\,g_{i''}\,g_{k''}$ but to $Q_{i'}\,Q_{k'}\,(g_{i''}-Q_{i''})(g_{k''}-Q_{k''})$. Equating this to the product $Q_{i''}\,Q_{k''}(g_{i'}-Q_{i'})(g_{k'}-Q_{k'})$, which determines the number of reverse transitions (the coefficients of proportionality being equal in both cases), we obtain the equilibrium condition in the form

$$\frac{Q_{i'}}{g_{i'}-Q_{i'}}\cdot\frac{Q_{k'}}{g_{k'}-Q_{k'}} = \frac{Q_{i''}}{g_{i''}-Q_{i''}}\cdot\frac{Q_{k''}}{g_{k''}-Q_{k''}}. \tag{138 b}$$

This shows, in conjunction with (138), that

$$\frac{Q_i}{g_i-Q_i} = Ae^{-\mu W_i}.$$

Replacing Q_i/g_i by $\overline{N_i}$ we get $\overline{N_i}/(1-\overline{N_i}) = Ae^{-\mu W_i}$, and therefore

$$\overline{N_i} = \frac{1}{\frac{1}{A}e^{\mu W_i}+1}.$$

Thus we again arrive at the Fermi-Pauli distribution law (135 b) derived in the preceding section for the same case but by a different method. A here replaces $e^{-\lambda}$ in (135 b).

To obtain the Bose-Einstein distribution law (134 b) we must obviously replace the negative signs in (138 b) by positive signs. Thus

for equilibrium we obtain

$$\frac{Q_{i'}}{g_{i'}+Q_{i'}}\cdot\frac{Q_{k'}}{g_{k'}+Q_{k'}}=\frac{Q_{i''}}{g_{i''}+Q_{i''}}\cdot\frac{Q_{k''}}{g_{k''}+Q_{k''}}, \tag{138 c}$$

which, in conjunction with the energy condition (138), gives

$$\frac{Q_i}{g_i+Q_i}=Ae^{-\mu W_i},$$

that is,

$$\overline{N_i}=\frac{1}{\frac{1}{A}e^{\mu W_i}-1},$$

in accordance with (134 b).

This result is easier to obtain than to interpret. For if equation (135 b) for the exclusive distribution is interpreted by admitting that the number of vacant states in a set i must be reduced by the number of particles Q_i belonging to the set, then we are forced to the conclusion that, for the inclusive distribution, the number of vacant states is, on the contrary, *increased* by each new intruder. This result has been expressed by L. Brillouin in a form which appears to be slightly less absurd. Thus if we replace the quantized states by the corresponding unit cells of the phase space and therefore replace the sets of states by volumes containing many adjacent cells, then we can say that particles described by antisymmetrical wave functions take, as it were, a positive volume of magnitude $\frac{1}{2}h^3$ and so reduce the space available for other particles, whereas particles described by symmetrical functions take a negative volume of the same magnitude and so increase the volume available for other particles. Of course this interpretation is in principle no better than the preceding one.

The difficulty is obviously due to the fact that in deriving the distribution law we have from the outset neglected the identity principle, i.e. we have disregarded the impossibility of identifying the 'different' particles forming our gas. We have reasoned as if we knew that *the very same particle* which is initially in some state of the set i' can be identified after the transition in some state of the set i''. Now such a reasoning is fundamentally wrong and so, if we wish to argue thus, we must be ready to accept the consequences. It would be a mistake to think that these consequences are more reasonable for exclusive than for inclusive distributions.

In both cases we are forced to ascribe almost anthropomorphic properties to the particles in order to interpret our fallacious conclusions concerning their considerate behaviour towards one another. It is just

as queer to imagine that an electron refrains from going into a state because the state is already occupied by another electron as to imagine that a photon is tempted to enter a state just because this state is already occupied by other photons. Such 'segregative' and 'herd' instincts can of course be ascribed to particles in a figurative sense only.

That the whole trouble is due to our trespassing against the identity principle can be illustrated by the following argument. If the particles are considered as distinguishable, then a distribution characterized by the partition numbers N_1, N_2,... is to be regarded as one of a number $g = \frac{N!}{N_1!\,N_2!\ldots}$ of equivalent but nevertheless different distributions. If, preserving the principle of individuality of the 'different' particles, we wish nevertheless to ascribe to all such distributions the same weight, e.g. the maximum weight $N!$, then we must assume that the number of ways in which N_i particles can be arranged *in the same state* is equal to $N_i!$. This would obviously mean that the probability of a new particle going into a given state would increase with the number N_i of particles already occupying the state, like $N_i!$. Replacing a state by a set of states we must replace the number $N_i!$ by

$$g_i(g_i+1)(g_i+2)\ldots(g_i+Q_i-1),$$

this product measuring the probability of Q_i particles being in a set of g_i states.

The preceding argument refers to an inclusive distribution. In the classical distribution the probability of finding Q_i particles in a set of g_i states is proportional to $g_i^{Q_i}$, whereas in exclusive distributions it is assumed to be measured by the product $g_i(g_i-1)\ldots(g_i-Q_i+1)$.

These results are closely related to the expressions (137) and (137a) derived above for the mean square deviations (fluctuations) from the equilibrium distribution. We shall not stop, however, to obtain the same expressions by the kinetic method discussed in this section.

29. Limiting Forms of the Quantum Distribution Laws

Boltzmann's classical distribution law can be regarded as the limiting form to which the two quantum distribution laws—that of Bose-Einstein expressed by (134b) and that of Pauli-Fermi-Dirac expressed by (135b)—tend with increase of the parameter λ. In fact for sufficiently large values of λ all three distribution laws assume the form

$$\overline{N_k} = e^{-\lambda-\mu w_k} \tag{139}$$

(if z is neglected compared with 1).

For numerically small values of λ (this parameter can be both positive and negative) the three distribution laws diverge widely from one another. The difference between them can be summed up as follows: The concentration of the particles in the states of lowest energy in the classical distribution is less than in the inclusive distribution and greater than in the exclusive distribution.

In practical applications, a system of particles, e.g. an ideal gas, is usually specified not by the parameter λ but by the number of particles constituting it, which can be identified with its average value $\overline{N}$. The specification is completed by the parameter μ, i.e. the temperature. The equation $\overline{N} = -\frac{\partial}{\partial\lambda}\log Z$ must then be used not to determine $\overline{N} = N$ but, conversely, to determine λ as a function of N and μ (or θ). This function is of course different for different types of distributions or 'statistics' and has in general a very complicated form. It can only be evaluated with ease when the temperature is very high or very low.

When the temperature is very high, i.e. when μ is very small, the particles must be distributed over a very *broad* range of states, extending from the normal state up to states for which the value $\mu W_k = W_k/\theta$, or rather the difference between this value and that corresponding to the normal state (W = minimum) becomes comparable with 1. Hence, unless the total number of particles N is extremely large, the average values $\overline{N_k}$ for the states lying within this range must be very small and consequently λ must be very large. The 'quantum' distribution laws, which can be represented by the single formula

$$\overline{N_k} = \frac{1}{e^{\lambda+\mu W_k} \pm 1}, \tag{139 a}$$

will thus reduce to the 'classical' form (139), so that λ will be determined approximately by the equation:

$$e^{-\lambda} \sum_k e^{-\mu W_k} = N,$$

or

$$\lambda = \log \frac{\sum e^{-\mu W_k}}{N}. \tag{140}$$

When, however, the temperature is very low, i.e. when μ is very large, λ can be defined approximately by the equation

$$\lambda = -\mu W_0, \tag{140 a}$$

where the subscript $_0$ refers to the state of lowest energy for an inclusive distribution, and to the state of highest energy to be found at the zero

point ($\mu = \infty$) for an exclusive distribution. In fact, substituting (140 a) in (139 a), we get

$$\overline{N_k} = \frac{1}{e^{\mu(W_k - W_0)} \pm 1}. \tag{140 b}$$

Considering first the lower sign, which corresponds to the concentration of all the particles in the lowest state, we see that $\overline{N_k} = \infty$ for $W_k = W_0$ and that $\overline{N_k} = 0$ (practically) for $W_k > W_0$. For the upper sign, however, which corresponds to the equipartition of the N particles among the N lowest states (cf. § 25), we find $\overline{N_k} = 1$ for $W_k < W_0$ and $\overline{N_k} = 0$ for $W_k > W_0$.

It may be remarked that λ tends to the form (140) not only when the temperature is very high but also when N (or $\overline{N}$) is comparatively small. The average values of $\overline{N_k}$ are small compared with 1 in both cases, so that for the inclusive and exclusive distributions we get practically the same equilibrium law (138) as for the classical distribution. When N becomes small the two quantum distribution laws become identical, but nevertheless remain different from the classical distribution law, inasmuch as they refer to a discrete series of quantized states instead of to the classical continuous series of states corresponding to $h \to 0$. This difference is particularly important at low temperatures and gradually vanishes as the temperature is raised. In fact for sufficiently small values of μ, and hence for sufficiently high temperatures, the product μW_k changes nearly continuously when we proceed from quantized states of lower energy to those of higher energy, i.e. μW_k changes *as if* the energy-steps were infinitely small, or as if the quantum constant h were infinitely small. But, as we have seen above, this is mathematically equivalent to replacing the quantum theory by the classical theory.

It should be emphasized once more that quantum statistics and classical statistics differ in two distinct ways—first in the definition of the individual states, and secondly with respect to the identity principle. The first of these distinctions is important when the successive quantized states lie far apart on the energy scale or when the temperature is low; the second is important when the number of particles is large.

We have just examined the case when N is small while the quantum energy-steps remain large. In this case we found that we could use the classical distribution law, but could not replace sums of the type $\sum_k e^{-\mu W_k}$ by integrals, as is done in the classical theory. We shall now examine the other case of a 'half-classical' distribution, namely, the

case of large values of N but small values of the energy-steps (multiplied by μ).

In this case we must preserve the form of the distribution law but can replace the quantized states by the corresponding volume elements of the classical phase space. If we make no distinction between the energy of the 'twin' quantum states associated with the two different values of the additional quantum number which specifies the spin of the electrons or the polarization of the photons, then the number of quantum states included in a phase volume $d\gamma = \int \dots \int dx \dots dg_z$ (for an individual particle) may be set equal to $2\,d\gamma/h^3$. The average number of particles which, in the equilibrium distribution, are in the classical 'state-range' $d\gamma$ can therefore be defined by the equation

$$dN = \frac{2}{h^3} \frac{d\gamma}{e^{\lambda+\mu W} \pm 1}, \tag{141}$$

the energy now being considered as a function of x, y, z, g_x, g_y, g_z. The two signs refer, as before, to the two types of quantum distribution. For an ideal gas enclosed in a vessel of volume V with no external forces acting on the particles we can put in (141)

$$d\gamma = V\, dg_x\, dg_y\, dg_z,$$

since in this case W depends only on the momentum components. Since W is a symmetrical function of these components, it is convenient to replace the 'rectangular' volume element of the 'momentum space' $dg_x\, dg_y\, dg_z$ by a spherical element $4\pi g^2\, dg$, which includes all values of the resultant momentum lying between g and $g+dg$, irrespective of its direction in space. This gives

$$d\gamma = 4\pi V g^2\, dg. \tag{141 a}$$

For the energy W we must use, in the general case, the relativistic expression

$$W = c\sqrt{\{(m_0 c)^2 + g^2\}},$$

or, if large velocities do not concern us, the ordinary non-relativistic formula

$$W = \frac{1}{2m} g^2.$$

(The rest-energy $m_0 c^2$ can be dropped as an immaterial constant.)

VI

APPLICATION OF THE QUANTUM STATISTICS TO THE ELECTRON THEORY OF METALS, HEAT MOTION AND RADIATION

30. The Theory of an Electron Gas; Free Electrons in Metals

The simplest application of the theory of an electron gas is to *metallic bodies*. These bodies are characterized by the presence of the so-called 'free' or 'conduction' electrons, which are responsible for their electrical conductivity and which—in the most typical cases at least—can be described with sufficient approximation as moving freely throughout the whole body, their mutual repulsion being on the average exactly balanced by the attraction which they experience on the part of the positive ions. This balance is of course destroyed at (or near) the surface of the body, where the electrons must experience a resulting attractive force pulling them inside. The 'freedom' of the conduction electrons has thus but a very relative meaning, being restricted to their motion within the metallic body, from which under ordinary conditions they are unable to escape.

The average balance of the forces acting on a 'free' electron within a metal can be also disturbed by the lack of regularity in the spatial distribution of the positive ions, due to their thermal agitation or to the presence of impurities. It seems natural, however, to assume that so long as these ions are distributed with a complete regularity, forming an ideal crystal lattice—a condition which may be found in absolutely pure metals only and at the temperature of the absolute zero—the *average* motion of the 'free' electrons will be the same as if they were acted on by no forces, except those forces, active in the surface layer, which serve to keep them inside the metal.

The metal can be treated accordingly as an empty box with completely reflecting walls, representing its surface. The surface can be visualized as a potential energy barrier of the simplified type discussed in § 15; the potential energy of an electron, crossing the surface S from the inside to the outside, is to increase by a certain amount U.

We shall consider first, on the basis of this simplified model of a metallic body, some of its properties which for ordinary temperatures are practically independent of the temperature, i.e. remain the same as at the temperature of absolute zero.

We shall assume that the number of 'free' electrons is of the same

order of magnitude as the number of atoms—as it should be if these electrons are recruited from the outer shell of each atom, thus corresponding to the 'valence' electrons of the atoms in the isolated state.†

It can easily be shown that under this condition the total *kinetic energy* of the electrons as well as its distribution among them is a property of the above kind, i.e. but very slightly affected by an increase of the temperature from $T = 0$ up to a few hundred and even a few thousand degrees.

The law representing the distribution of velocities or momenta in a gas constituted by a very large number of free electrons is expressed with sufficient accuracy by equation (141) taken with the *plus* sign.

Putting $\lambda = -\mu W_0$ and using the expression (141 a) for $d\gamma$, we can rewrite it in the form

$$dN = \frac{8\pi V}{h^3} \frac{g^2\, dg}{e^{\mu(W-W_0)}+1} \qquad \left(\mu = \frac{1}{\theta}\right). \tag{142}$$

As shown above, W_0 can be identified approximately with the maximum kinetic energy of an electron at the absolute zero of temperature, this approximation being satisfactory so long as θ is small compared with W_0. To ascertain the range of temperature for which this condition is fulfilled we must first determine W_0 for $\theta = 0$. Since $e^{\mu(W-W_0)}$ is then equal to 0 for $W < W_0$ and to ∞ for $W > W_0$, the expression (142) reduces to

$$dN = \begin{cases} \dfrac{8\pi V}{h^3}\, g^2\, dg & (g \leqslant g_0) \\ 0 & (g > g_0), \end{cases}$$

where g_0 is the maximum value of the momentum of an electron at $\theta = 0$. It is related to W_0 by the equation $W_0 = g_0^2/2m$. We thus get

$$N = \frac{8\pi V}{h^3} \int_0^{g_0} g^2\, dg = \frac{8\pi V}{3h^3}\, g_0^3,$$

or

$$n = \frac{N}{V} = \frac{8\pi}{3h^3} g_0^3. \tag{142 a}$$

Hence the maximum zero-point energy of an electron is expressed in terms of the number n of electrons per unit volume by the formula

$$W_0 = \frac{h^2}{2m}\left(\frac{3n}{8\pi}\right)^{\frac{2}{3}}. \tag{142 b}$$

It should be noticed that the quantity $h/g_0 = \lambda_0$ is just the *shortest wave-length* of the free electrons in the metal, while $1/n^{\frac{1}{3}}$ would be the

† See below, § 31.

distance between neighbouring electrons if they were all arranged into a cubic lattice. Thus equation (142a) becomes

$$\lambda_0 = \left(\frac{8\pi}{3n}\right)^{\frac{1}{3}}. \tag{142c}$$

Thus the shortest wave-length of the free electrons at the zero-point of temperature is of the same order of magnitude as the distance between neighbouring atoms, since the number of atoms is equal to or comparable with the number of free electrons. Now the distance between neighbouring atoms in a solid body is of the same order of magnitude as the linear dimensions of these atoms. These linear dimensions are approximately equal to the diameter or the length of the orbit of the external electrons in an isolated atom. According to de Broglie's relation the length of this orbit is of the same order of magnitude as the wave-length of these electrons. We thus see that the wave-length and consequently the velocity of the fastest free electrons in a metallic body must be approximately equal to the wave-length and velocity of the 'valency' electrons in isolated metallic atoms. This velocity is, as is well known, of the order of magnitude of 10^8 cm./sec. and corresponds to an ionization energy (or to an average kinetic energy) of the order of magnitude of a few volts, while the thermal energy, as defined by the classical formula $\frac{3}{2}kT$, is, for ordinary temperatures, about 100 times smaller. We thus see that the energy W_0 is approximately 100 times larger than the quantity $\theta = kT$ for ordinary temperatures, and that even for temperatures of a few thousand degrees the quantity W_0 in equation (142) can be treated as a constant and identified with the 'zero-point maximum energy' (142b).

The electron gas under this condition is called *degenerate*, because, according to equation (142), its energy is practically independent of the temperature, and its heat capacity is very small compared with the classical value $\frac{3}{2}k$ per electron. The average value of the energy W per electron can be calculated by means of the equation

$$\overline{W} = \frac{E}{N} = \frac{1}{2m}\int_0^\infty \frac{g^4\,dg}{e^{\mu(W-W_0)}+1} \Bigg/ \int_0^\infty \frac{g^2\,dg}{e^{\mu(W-W_0)}+1}. \tag{143}$$

At $T = 0$ it is equal to

$$\overline{W} = \tfrac{3}{5}W_0. \tag{143a}$$

Its dependence upon the temperature or rather the classical heat energy $\frac{3}{2}kT$ is shown in Fig. 16. For sufficiently high temperatures W tends asymptotically to the classical value $\frac{3}{2}kT$. At lower temperatures its

increase is the slower the larger its initial value. Its dependence upon the temperature in this region can easily be determined in the following approximate manner. For temperatures satisfying the condition $kT \ll W_0$, the expression $f = 1/(e^{\mu(W-W_0)}+1)$ considered as a function of W is practically a straight line $f = 1$ so long as W is markedly smaller than W_0. As W approaches this critical value, $f(W)$ begins to drop, becomes equal to $\frac{1}{2}$ at $W = W_0$, and tends asymptotically to zero as W increases beyond this value (Fig. 17). The two areas shown shaded in the figure are equal, for they represent the number ΔN of electrons whose energies (lying close to W_0) are shifted from $W < W_0$ to $W > W_0$

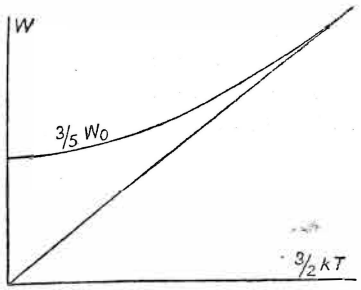

FIG. 16

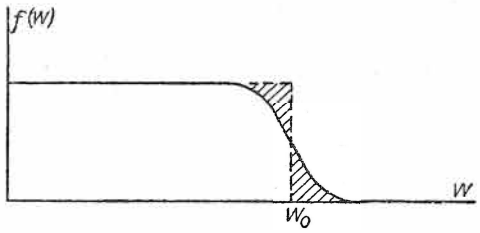

FIG. 17

as the temperature increases from zero to the given value. In fact (142), in conjunction with $g\,dg/m = dW$, gives

$$\Delta N = \frac{8\pi Vm}{h^3}\int_{W_0}^{\infty}\frac{g\,dW}{e^{\mu(W-W_0)}+1} = \frac{8\pi Vm}{h^3}\int_{0}^{W_0} g\,dW\left(1-\frac{1}{e^{\mu(W-W_0)}+1}\right) \quad (144)$$

if the total number of electrons N is to remain constant. We may replace g in the integrands by g_0, the difference $g-g_0$ being sensible only for vanishing values of the integrands. On putting $W-W_0 = \xi$ we therefore require

$$\int_0^{\infty}\frac{d\xi}{e^{\mu\xi}+1} = \int_{-\infty}^{0}\left(1-\frac{1}{e^{\mu\xi}+1}\right)d\xi, \quad (144\,a)$$

which is true and justifies the above statement. For, putting $\xi = -\xi'$ in the second integral, we have

$$1-\frac{1}{e^{\mu\xi}+1} = \frac{e^{\mu\xi}}{e^{\mu\xi}+1} = \frac{1}{1+e^{-\mu\xi}} = \frac{1}{e^{\mu\xi'}+1}$$

so that $$\int_{-\infty}^{0}\left(1-\frac{1}{e^{\mu\xi}+1}\right)d\xi = -\int_{\infty}^{0}\frac{1}{e^{\mu\xi'}+1}\,d\xi' = \int_0^{\infty}\frac{d\xi}{e^{\mu\xi}+1}.$$

When the temperature is raised the increase ΔE in the total energy E of the electron gas above its value at the absolute zero can be represented by the formula

$$\Delta E = \frac{8\pi Vm}{h^3}\Big[\int_0^{W_0} Wg(f-1)\,dW + \int_{W_0}^{\infty} Wgf\,dW\Big],$$

or approximately

$$\Delta E \cong \frac{8\pi Vmg_0}{h^3}\Big[\int_{-\infty}^{0} (W_0+\xi)(f-1)\,d\,\xi + \int_0^{\infty} (W_0+\xi)f\,d\xi\Big].$$

In virtue of (144 a) this reduces to

$$\Delta E = \frac{16\pi Vmg_0}{h^3}\int_0^{\infty} \frac{\xi\,d\xi}{e^{\mu\xi}+1} = \frac{16\pi Vmg_0}{h^3\mu^2}\int_0^{\infty} \frac{\eta\,d\eta}{e^{\eta}+1} \qquad (\eta = \mu\xi).$$

The integral $I = \int_0^{\infty} \frac{\eta\,d\eta}{e^{\eta}+1}$ can be easily evaluated by expanding the denominator according to the formula

$$1/(e^{\eta}+1) = e^{-\eta}/(1+e^{-\eta}) = e^{-\eta}\sum_{n=0}^{\infty}(-1)^n e^{-n\eta}.$$

This gives

$$I = \sum_{n=0}^{\infty}\frac{(-1)^n}{(n+1)^2} = 1-\frac{1}{2^2}+\frac{1}{3^2}-\frac{1}{4^2}+\ldots = \frac{\pi^2}{12}.$$

Dividing the preceding expression for ΔE by N and using (142 a) and the relation $g_0^2 = 2mW_0$, we get

$$\Delta\overline{W} = \frac{\pi^2}{4W_0}k^2T^2. \tag{145}$$

Hence the heat capacity per electron of the electron gas is equal to

$$C_e = \frac{\pi^2k^2}{2W_0}T. \tag{145 a}$$

It vanishes at $T = 0$, and is extremely small compared with the classical value $C = \frac{3}{2}k$ at ordinary temperatures (in fact the ratio

$$C_e/C = \tfrac{1}{3}\pi^2kT/W_0$$

is of the order of magnitude of $\frac{1}{100}$ when $T \cong 273$).

It should be remembered that the preceding results are only valid if the 'energy steps' between successive quantized states are very small compared with kT. It is only in this case that sums over such states can be replaced by integrals over the phase space. It can easily be shown that this condition is fully satisfied not only for ordinary but even for very low temperatures. A rough estimate of the value of the

energy steps is obtained by dividing the maximum energy W_0 by the total number of steps which is equal to half the number of electrons N. Since W_0 does not depend upon N but only upon the *concentration* of the electrons $n = N/V$, we see that the energy steps of an electron gas are the smaller the larger the volume V of the metallic body 'containing' the electron gas. The ratio of $2W_0/N$ to kT, even when $W_0/kT = 1{,}000$, must be extremely small for metallic bodies of ordinary size, since the number of electrons N is of the same order of magnitude as the number of atoms. Even in the smallest metallic crystals N is of the order of magnitude of 10^{10}.

In the foregoing considerations we have assumed that 'twin' states, differing from one another only by the orientation of the electrons, have the same energy. This equality is destroyed by the presence of an *external magnetic field.* If the field strength is **H**, and if we assume that the magnetic moments **M** of the electrons are oriented either parallel or anti-parallel to the field, then the additional magnetic energies of the twin states have different values, namely, $-MH$ in the parallel and $+MH$ in the anti-parallel case. We shall now examine the resultant magnetic moment of the metallic body under these conditions *at the zero-point of temperature.*

When no magnetic field is present the N electrons are distributed in oppositely oriented pairs among the $s = \frac{1}{2}N$ lowest stationary states or 'cells' of the phase-space skyscraper. When a field is applied it tends to make all the electrons turn in the direction of the field. Since, however, two electrons cannot remain in the same cell while pointing in the same direction, they must part company if both are to be oriented by the field, and one of them must move to a vacant cell of higher energy. Thus the decrease of energy due to the orientation of a pair of electrons in the direction of the magnetic field is at least partially compensated by the increase of kinetic energy due to the transfer of one of them to a higher floor of the phase-space skyscraper. Before leaving its partner and moving upstairs, an electron must therefore decide, as it were, whether it pays to do so or not, i.e. whether this will mean a net decrease or a net increase of energy. In the former case it will move, for the normal state of a system of electrons at the zero-point of temperature is that of least possible energy consistent with the exclusion principle; in the latter case it will stay where it is.

As the field is increased, more and more electrons will be induced to move to higher states. In the absence of the magnetic field the highest cell was the $\frac{1}{2}N$th one. Let N' be the number of electrons that have

moved upstairs when the field strength is H. The $2N'$ highest cells will each be occupied by a single electron, the $(\frac{1}{2}N+N')$th being the highest occupied cell. As the field is increased the next pair of electrons to be oriented will be those in the uppermost doubly occupied cell on the $(\frac{1}{2}N-N'-1)$th floor. One of the two electrons will remain in this cell, while the other will move to the lowest vacant cell, which is situated on the $(\frac{1}{2}N+N'+1)$th floor. If N' is very small compared with N, then the height between successive floors, i.e. the difference of kinetic energy ΔW between successive states, is approximately constant. Therefore the height through which one of the electrons will have to be lifted, i.e. the increase of its kinetic energy, is equal to $2(N'+1)\,\Delta W$. If the N'th electron is the last to move upstairs, we must have

$$2N'\,\Delta W < 2MH \quad \text{and} \quad 2(N'+1)\,\Delta W > 2MH$$

or approximately, so long as N' is very large compared with 1,

$$N' = \frac{MH}{\Delta W}. \tag{146}$$

The total magnetic moment due to the orientation of $2N'$ electrons is equal to $2N'M = \chi H$, where $\chi = \dfrac{2M^2}{\Delta W}$. χ is the *magnetic susceptibility* of the electron gas.

Since ΔW is of the order of magnitude of $2W_0/N$, χ must be of the order of magnitude of M^2N/W_0.

We can determine ΔW more exactly from the relation

$$4\pi V g_0^2\,\Delta g = h^3$$

for the volume of the $\frac{1}{2}N$th cell, in conjunction with the relation

$$\Delta W = \frac{1}{m} g_0\,\Delta g.$$

This gives $$\Delta W = \frac{h^3}{4\pi m V g_0} = \frac{h^3 g_0^2}{4\pi m V g_0^3} = \frac{h^3 W_0}{2\pi V g_0^3}$$

and hence, by (142 a), $$\Delta W = \frac{4W_0}{3N},$$

so that finally $$\chi = \frac{3}{2}\frac{M^2N}{W_0}. \tag{146 a}$$

This formula was derived by Pauli in 1927. It determines the positive or paramagnetic part of the magnetic susceptibility of a metal, i.e. the part depending upon the orientation of the electron spins. It remains valid for temperatures which are sufficiently small compared with W_0/k and for magnetic fields which are sufficiently weak compared with

W_0/M. With $M \cong 10^{-20}$ (Bohr magneton) and $W_0 \cong 10^{-14}$ erg (this corresponds to about 5 volts) we get $W_0/M \cong 10^5$ gauss, which is much higher than the fields usually employed. Thus in the region of ordinary field strengths and ordinary temperatures the 'spin paramagnetic susceptibility' of a metal is independent both of the field strength and of the temperature.

Leaving magnetic effects aside, let us consider now two other properties of a metallic body, which can be derived from the electron gas theory, and which are but slightly affected by the temperature. These properties are the *density* and the *compressibility*.

If we assume that the number of free electrons is equal to the number of atoms (an assumption fully justified in the case of alkali metals), the density of the body can be calculated from the concentration of the free electrons. In the preceding considerations this concentration was assumed to be known and the mean kinetic energy of the electrons (at $T = 0$) was calculated therefrom with the help of (142 b) and (143 a). This connexion between the concentration of the free electrons and their mean kinetic energy seems at first sight to be wholly devoid of any dynamical foundation, being merely a corollary of the Pauli exclusion principle. As a matter of fact, however, the motion of the electrons in a metal—just as in a simple atom—must be closely related to the *forces* acting upon them, larger velocities corresponding to larger forces.

This relation can be put into a more precise form by introducing instead of the forces the corresponding potential energy U. In ionic crystals, such as rock salt, the potential energy of an ion with respect to the rest, so far as it depends upon their electric charges, is equal to $-\gamma e^2/R$, where γ is a numerical constant of the order of magnitude 1, $\pm e$ the charge of the ions, and R the distance between the neighbouring ions. This formula can be used for the approximate determination of the potential energy of the free electrons in a metallic body, since on the average they must be found between the positive ions (binding them together in the same way as the Cl^- ions bind the Na^+ ions in rock salt). Replacing $1/R$ by $n^{\frac{1}{3}}$ (which means a slight readjustment of the numerical factor γ) and multiplying by the total number of (free) electrons N, we get the following expression for the potential energy of the electrons:

$$NU = -\gamma e^2 n^{\frac{1}{3}} N.$$

Adding to this their kinetic energy

$$N\overline{W} = \tfrac{3}{5} N W_0 = \frac{3}{10}\frac{h^2}{m}\left(\frac{3n}{8\pi}\right)^{\frac{2}{3}} N,$$

we obtain the total energy E as a function of the electron concentration

$$E = \alpha n^{\frac{2}{3}} - \beta n^{\frac{1}{3}}, \tag{147}$$

with $$\alpha = \frac{3}{10}\frac{h^2}{m}\left(\frac{3}{8\pi}\right)^{\frac{2}{3}} N, \qquad \beta = \gamma e^2 N.$$

This formula enables one to calculate the actual value of n in the equilibrium condition which can be defined as corresponding to the minimum of the total energy of the electrons. We have in fact $\partial E/\partial n = 0$, that is, according to (147),

$$\tfrac{2}{3}\alpha n^{-\frac{1}{3}} - \tfrac{1}{3}\beta n^{-\frac{2}{3}} = 0,$$

or

$$n^{\frac{1}{3}} = \tfrac{1}{2}\beta/\alpha. \tag{147 a}$$

It should be noticed that this formula corresponds to the following relation between the kinetic and potential energy

$$2\overline{W} = -U, \tag{147 b}$$

which can be obtained directly from the so-called 'virial theorem' for the special case of a system of particles moving under the action of electrical forces (see Part II).

Introducing in (147 a) the numerical values of the constants α and β, we find

$$n^{\frac{1}{3}} \cong 5\gamma \,.\, 10^8 \text{ cm.}^{-1},$$

which corresponds to a distance of the correct order of magnitude (~ 1 Å U.) between neighbouring atoms.

We thus see that the kinetic energy of the electron gas, as determined on the basis of Pauli's exclusion principle, may be put in complete harmony with the value resulting from purely dynamical principles by the proper choice of the electron concentration n.

If the metal is slightly compressed or extended by some external force in the form of positive or negative pressure, its energy E is increased with respect to the equilibrium value by the amount

$$E - E_0 = \frac{1}{2}\left(\frac{\partial^2 E}{\partial R^2}\right)_0 (R - R_0)^2.$$

Putting $R = n^{-\frac{1}{3}} = \left(\frac{V}{N}\right)^{\frac{1}{3}}$, we have $R - R_0 = \frac{1}{3}\left(\frac{V_0}{N}\right)^{\frac{1}{3}} \frac{V - V_0}{V_0}$, and consequently

$$E - E_0 = \tfrac{1}{2}\kappa \frac{(V - V_0)^2}{V_0}, \tag{148}$$

where

$$\kappa = \frac{1}{9}\frac{R_0^2}{V_0}\left(\frac{d^2 E}{dR^2}\right)_0. \tag{148 a}$$

The quantity $E - E_0$ can be defined as the *elastic* energy of the body. Its differential coefficient with regard to the volume, $\kappa(V - V_0)/V_0$, must

be equal and opposite to the external pressure applied to the body. We thus see that κ represents the *modulus of compressibility* of the body. Using the expression (147) for the energy, and taking into account the relation (147 a), we easily get

$$\kappa = \frac{1}{9V_0}\frac{\beta}{R_0} = \frac{\gamma}{9}e^2n^{\frac{4}{3}}. \tag{148 b}$$

This formula is in satisfactory agreement with the experimental data on the compressibilities of the alkali metals. The agreement is of course only roughly approximate in view of the simplifications which have been made in the derivation of the expression (147) for the energy E.

The potential energy U referred to a single electron is equal and opposite to the work which must be done against the forces acting in the surface-layer of the metal on an electron escaping from it. To obtain the actual work which is required for this we must subtract from $|U|$ the kinetic energy of the electron W while it is moving within the metal. At the zero point of temperature the minimum value of this work is equal to

$$w = |U| - W_0, \tag{149}$$

where W_0 is the maximum kinetic energy. Now according to (147 b) and (143 a) we have

$$|U| = \tfrac{6}{5}W_0, \tag{149 a}$$

whence

$$w = \tfrac{1}{5}W_0. \tag{149 b}$$

The magnitude of w can be determined experimentally from the photo-electric threshold of metals (i.e. from the minimum frequency $\nu_0 = w/h$ which is required for the extraction of an electron by the action of light) or from the 'work function' of the Richardson effect (see below). It turns out to be of the order of magnitude of a few volts (2–3 volts for alkali metals). This corresponds, according to (149 b), to values of W_0 which are about six times larger than those calculated by means of the formula (142 b). The relation $|U| = 6w$ obtained from (149 a) and (149 b) with the experimental values of w leads, however, to values of $|U|$ which—at least in the case of the alkali metals—are in fair agreement with those obtained experimentally from the refractive index of metals for cathode rays (see Ch. II).

These results are supported by the following argument, based on the application of the virial theorem. The latter can be put in the form

$$\overline{W} = -(\overline{W} + U) = -\overline{E}_1,$$

where E_1 is the total energy referred to a single electron. In the case of an isolated alkali atom, Na say, the absolute magnitude of this

energy is equal to the work required to remove the valence electron, i.e. to the *ionization energy* of the atom, 5·1 electron-volts.

When a number of isolated atoms, forming the vapour of the metal, condense into a liquid or solid body, the total energy is diminished by the *latent heat of condensation*. Now since condensation is due to the action of cohesive forces, which depend upon the mutual action of the electrons and the positive ions, and since, further, the energy of heat motion of the latter is relatively very small, the latent heat of condensation can be identified with the decrease of the total energy of the electrons, which in the process of condensation change their role from that of the 'valence electrons' in the separate atoms to that of 'free electrons' in the metal. According to the virial theorem this decrease of E must be accompanied by a numerically equal increase of the average kinetic energy.

The heat of condensation of sodium amounts to 26,000 calories per gram atom, which is equivalent to 1·1 volts per electron. We thus see that in the solid or liquid state of this substance the average kinetic energy of a 'free' electron must be larger than in an isolated atom by the amount 1·1. Adding this to the ionization potential (5·1) we get 6·2 volts—whereas the formula

$$\overline{W} = \tfrac{3}{5}W_0 = \frac{3}{10}\frac{h^2}{m}\left(\frac{3n}{8\pi}\right)^{\frac{2}{3}}$$

of the Pauli-Sommerfeld theory gives for sodium ($n = 2{\cdot}5.10^{22}$) $\overline{W} = 2{\cdot}4.10^{-12}$ ergs or 1·5 volts. Similar results are obtained for the other alkali metals. Although the absolute value of the average kinetic energy given by the preceding formula is erroneous, yet the relation between $\overline{W}$ and W_0 which depends upon the distribution of velocities between the electrons seems to be correct, for it leads in conjunction with (149 b) to values of w which are in close agreement with the experimental data—as shown by the following table:†

The work functions of various metals.

Metal	Na	K	Rb	Cs	Cu	Ag	Zn	Cd	Hg	Ca
(calc.)	2·1	1·8	1·7	1·6	3·8	3·6	4·7	4·5	5·2	3·1
(exper.)	1·8	1·8	1·7	1·6	4·0	3·8	3·2	3·6	4·0	3·3

Multiplied by 6 these values should give, according to (149 a) and (149 b), the potential energy $|U|$. As a matter of fact the values so obtained turn out to be of the correct order of magnitude (when com-

† Taken from Frenkel, *Zeits. f. Physik.* **49**, p. 40 (1928). The 'calculated' values of w (in volts) are derived from the ionization potentials and the condensation heat of the corresponding metals.

pared with cathode rays diffraction data) for the alkali metals only, and nearly twice too large for other metals (for instance, 22·8 for Ag, instead of 14, according to Rupp's data).

31. Refinement and Extension of the Electron Gas Theory

The failure of the Pauli-Sommerfeld electron gas theory to give correct numerical values for w and the related constants of metals must obviously be traced to the unwarranted simplification of the motion of the 'free' electrons. It is certainly incorrect to identify this motion with that in force-free space. This can be seen most clearly by starting with the motion of the external or valence electrons in the separate atoms of a metallic vapour (sodium, say), and considering the gradual transformation of this motion during the process of condensation into a solid body. For the sake of simplicity this process can be pictured as an 'adiabatic' one, i.e. as an infinitely slow approach of the atoms until their nuclei form the ideal crystalline lattice, characteristic of the equilibrium state of the metal at the zero point of temperature.

We shall first consider the motion of the electrons from the point of view of classical mechanics.† In the case of the isolated sodium atoms it reduces to a revolution about the one nucleus or ion. When the atoms are brought close together at a distance of the same order of magnitude as the 'orbits' of the external electrons in the separate atoms, the rather loosely attached electrons no longer remain bound to the same atoms, but begin to move about from one atom to the next throughout the whole body. They thus become free from individual serfdom but remain bound to the body as a whole, and indeed *on the average* even more firmly bound than before (the increase of the binding energy being equal to the latent heat of condensation referred to one electron).

So long as the atoms (ions) are arranged in a perfectly regular crystalline lattice, and the mutual action of the electrons is replaced by an equivalent ('self-consistent') external field, the motion of each electron can be visualized as similar to the motion of a point on the circumference of a wheel (or the surface of a ball) rolling on the average uniformly in a straight line. It can of course stop at each atom for a while, allowing the electron to perform a certain number (s) of revolutions about its nucleus and then pass to the next, where the same story is repeated once more, and so on. If the period of revolution is equal to τ and the distance between neighbouring atoms in the line of 'rolling'

† Following the lines of an argument presented in *Zeits. f. Physik*, **29**, 214 (1924).

to d, then the average velocity with which the electron will travel through the metallic body can be defined by the equation

$$v = \frac{d}{s\tau}.$$

Even in the most favourable case (of one revolution per atom) it must be smaller than the true velocity, defined by the kinetic energy of the combined motion. Now in the theory of Pauli and Sommerfeld the periodic part of this motion is wholly neglected, the 'electron-wheels' being replaced by their 'centres' and the motion of the free electrons visualized as the motion of these centres. No wonder, then, that this theory leads to too small values of the average kinetic energy and of the associated constants.

It should be mentioned that the rectilinear uniform motion (rolling) of the electron-wheels can take place only because the lattice remains perfectly regular. Any lack of regularity, due to the presence of impurities or to the heat motion of the ions, will cause the electron-wheels to deviate from the straight line, the rectilinear segments laid down between two such successive deviations playing the same role as the 'mean free path' in the case of the atoms of an ordinary gas. We shall come to this question later on in connexion with the electrical conductivity of metals.

The above classical or corpuscular picture of the motion of the free electrons in a crystalline lattice is in full harmony with the wave-mechanical picture, corresponding to it in precisely the same way as the classical picture of the orbital motion of the valence electron about an isolated atom corresponds to the wave-mechanical picture discussed in § 14, in connexion with the Schrödinger theory of a hydrogen-like atom. An essential advantage of the wave-mechanical theory is that it lends itself much better to the simplification of replacing the mutual action of the various electrons by the action of an 'equivalent' or 'self-consistent' external electric field—this replacement leading to much less error with wave mechanics than with classical mechanics.

Assuming this error to be small enough, we can determine the motion of a free electron wave-mechanically with the help of the ordinary Schrödinger equation (for a single electron)

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(E-U)\psi = 0, \tag{150}$$

where the potential energy U must be suitably defined *as a periodic*

function of the coordinates x, y, z with the same periodicity as that of the crystal lattice under consideration.

This problem has been considered already in § 16 for the simplest case of a crude one-dimensional model of a crystal lattice. The main results obtained there can easily be extended to the general case. In particular it can easily be shown that the solution of our problem—i.e. the motion of the electron in the periodic field of force of a three-dimensional lattice—is represented by a wave function of the type

$$\psi = C(x,y,z)e^{i(\gamma_x x+\gamma_y y+\gamma_z z)} = C(\mathbf{r})e^{i\boldsymbol{\gamma}\cdot\mathbf{r}}, \tag{150a}$$

where C is a certain periodic function of the coordinates with the same periodicity as the potential energy $U(x,y,z)$ and consequently as the lattice itself. So long as the parameters γ_x, γ_y, γ_z remain all real, the function (150) represents just that kind of combined motion which has been described above, the factor $e^{i\boldsymbol{\gamma}\cdot\mathbf{r}}$ corresponding to the uniform rectilinear motion of the centre of the 'electron-wheel' and the factor $C(\mathbf{r})$ to its revolution. With an infinite crystal lattice the values of the parameters γ_x, γ_y, γ_z can vary in a continuous manner, being limited only by the condition that the energy E which is a certain function of these parameters of the type

$$E = F(\gamma_x,\gamma_y,\gamma_z) \tag{150b}$$

should have *real* characteristic values. The values of $\boldsymbol{\gamma}$ can, however, in general be both real and imaginary, the latter case corresponding to selective reflection.

We shall not stop here to make a thorough investigation of the functions (150a) and (150b), but shall limit ourselves to their approximate determination for relatively large values of E, corresponding to a positive value of the kinetic energy of the electron $E-U$ throughout the lattice.

The potential energy can be represented as a periodic function of the coordinates by a Fourier series,

$$U = U_0 + \sum_{q_1,q_2,q_3}{}' U_{\mathbf{q}}\, e^{i\mathbf{q}\cdot\mathbf{r}}, \tag{151}$$

where U_0 is the average value of the potential energy and $\mathbf{q}$ is a vector with the components $q_s = 2\pi n_s/c_s$ ($s = 1,2,3$; $n_s = 0, \pm 1, \pm 2, \ldots$), if we consider a rectangular lattice with the constants (spacings) c_1, c_2, c_3. The accent on $\sum$ in (151) denotes the exclusion of the constant term

$$n_1 = n_2 = n_3 = 0.$$

If the expression (151) is introduced in Schrödinger's equation, and if the periodic terms of U are dropped as small compared with the

average kinetic energy of the electron $E-U_0 = W$, we obtain an approximate solution of the type $\psi^0 = e^{i\boldsymbol{\gamma}\cdot\mathbf{r}}$ with

$$\gamma = \frac{2\pi}{\lambda} = \frac{2\pi mv}{h} = \frac{2\pi}{h}\sqrt{(2mW)}.$$

This solution represents the 'primary' wave travelling through the crystal in an arbitrarily chosen direction. To every periodic term of U there corresponds an 'associated' wave, which to a first approximation is determined by the equation

$$\nabla^2\psi'_{\mathbf{q}}+\gamma^2\psi'_{\mathbf{q}} = V_{\mathbf{q}}\, e^{i\mathbf{q}\cdot\mathbf{r}}\psi^0 = V_{\mathbf{q}}\, e^{i(\mathbf{q}+\boldsymbol{\gamma})\cdot\mathbf{r}}, \tag{151 a}$$

obtained from (150) by putting $\psi = \psi^0+\sum_{\mathbf{q}}{}' \psi'_{\mathbf{q}}$ and dropping small terms of the second order, i.e. products of the form $V_{\mathbf{q}}\psi'_{\mathbf{p}}$, where

$$V = 8\pi^2 mU/h^2.$$

The solution of (151 a) is given by

$$\psi'_{\mathbf{q}} = A_{\mathbf{q}}\, e^{i(\mathbf{q}+\boldsymbol{\gamma})\cdot\mathbf{r}}, \tag{151 b}$$

with

$$A_{\mathbf{q}} = \frac{V_{\mathbf{q}}}{(\mathbf{q}+\boldsymbol{\gamma})^2-\gamma^2} = \frac{V_{\mathbf{q}}}{\mathbf{q}^2+2\mathbf{q}\cdot\boldsymbol{\gamma}}. \tag{151 c}$$

We thus get the following expression for the function ψ:

$$\psi = \psi^0+\sum{}' \psi'_{\mathbf{q}} = C_{\boldsymbol{\gamma}}(\mathbf{r})e^{i\boldsymbol{\gamma}\cdot\mathbf{r}}, \tag{152}$$

where $C_{\boldsymbol{\gamma}}(\mathbf{r})$ is a periodic function with the same type of periodicity as that characteristic of the structure of the lattice, namely

$$C_{\boldsymbol{\gamma}}(\mathbf{r}) = 1+\sum_{\mathbf{q}}{}' \frac{V_{\mathbf{q}}\, e^{i\mathbf{q}\cdot\mathbf{r}}}{\mathbf{q}^2+2\mathbf{q}\cdot\boldsymbol{\gamma}}, \tag{152 a}$$

in agreement with the general formula (150 a).

It should be noticed that our approximation is valid only so long as all the coefficients $A_{\mathbf{q}}$ are small, that is, all the denominators $\mathbf{q}^2+2\mathbf{q}\cdot\boldsymbol{\gamma}$ are large, or at least non-vanishing. The physical meaning of this condition can be seen as follows: Let us introduce a quantity $2\pi/q$ which is connected with the vector $\mathbf{q}$ in the same way as the wave-length λ is connected with the vector $\boldsymbol{\gamma}$ ($\lambda = 2\pi/\gamma$). This quantity must obviously be equal to a sub-multiple of the spacing d between the successive crystalline planes perpendicular to $\mathbf{q}$. We thus have $q = 2\pi n/d$ (n = integer). Introducing further the angle θ between the vectors $\boldsymbol{\gamma}$ and $-\mathbf{q}$, we have

$$\mathbf{q}^2+2\mathbf{q}\cdot\boldsymbol{\gamma} = 4\pi^2\frac{n}{d}\left(\frac{n}{d}-\frac{2}{\lambda}\cos\theta\right),$$

whence it follows that formula (152) ceases to hold for values of θ which satisfy the relation

$$2d\cos\theta = n\lambda. \tag{152 b}$$

This is Bragg's well-known relation which determines the condition for selective reflection and which has been discussed already in § 16 for the special case $\theta = 0$.

So long as it is not fulfilled, the expression (152) for ψ remains valid and the kinetic energy of the electron is given with sufficient accuracy by the 'zero order' expression

$$W = \frac{h^2\gamma^2}{8\pi^2 m} = \frac{h^2}{8\pi^2 m}(\gamma_x^2+\gamma_y^2+\gamma_z^2).$$

This corresponds obviously to the Pauli-Sommerfeld theory of the motion of the free electrons in a metal, a motion which is practically unaffected by the periodic character of the field of force of the metallic crystal lattice.

We shall now briefly consider the opposite extreme case of an electron moving with a kinetic energy which is relatively small and even highly negative in the inter-atomic regions. For the sake of simplicity we shall limit ourselves to the one-dimensional case investigated in § 16. The energy E is here connected through the parameters

$$\alpha = \frac{2\pi}{h}\sqrt{2m(E-U_a)}, \qquad \beta = \frac{2\pi}{h}\sqrt{2m(U_b-E)}$$

with γ by the relation

$$\cos\gamma c = \frac{\beta^2-\alpha^2}{2\alpha\beta}\sin\alpha a \sinh\beta b + \cos\alpha a \cosh\beta b.$$

For large values of βb, the allowed values of the energy, corresponding to real values of γ are confined to the neighbourhood of the energy-levels

$$E_0 = \frac{h^2 n^2}{8ma^2} + U_0$$

corresponding to the stationary states of the electron enclosed between two impermeable walls ($\beta b = \infty$) at a distance $a = 2l$ from each other.

Now we have for such states, according to (78 a) and (78 b), § 16, one of the two relations

$$\tan\tfrac{1}{2}\alpha_0 a = \beta_0/\alpha_0 \quad \text{or} \quad \cot\tfrac{1}{2}\alpha_0 a = -\beta_0/\alpha_0.$$

Taking into account the identities

$$\sin x = 2\sin\tfrac{1}{2}x\cos\tfrac{1}{2}x = \frac{2\tan\frac{1}{2}x}{1+\tan^2\frac{1}{2}x}; \quad \cos x = 1-2\sin^2\tfrac{1}{2}x = \frac{1-\tan^2\frac{1}{2}x}{1+\tan^2\frac{1}{2}x},$$

we thus get

$$\sin\alpha_0 a = \pm\frac{2\beta_0\alpha_0}{\alpha_0^2+\beta_0^2}, \qquad \cos\alpha_0 a = \frac{\alpha_0^2-\beta_0^2}{\alpha_0^2+\beta_0^2}.$$

The *plus* sign corresponds to the symmetrical case, and the *minus* sign to the anti-symmetrical one.

Substituting these expressions into the preceding general expression for $\cos\gamma c$ and replacing $\sinh\beta b$ and $\cosh\beta b$ by $\frac{1}{2}e^{\beta_0 b}$, we get for values of α and β lying in the neighbourhood of α_0 and β_0

$$\cos\gamma c = \frac{e^{\beta_0 b}}{2(\alpha_0^2+\beta_0^2)}[\pm(\beta^2-\alpha^2)+\alpha_0^2-\beta_0^2].$$

It is easily seen that the minus sign must be dropped, for with $\beta_0 b \gg 1$ it leads to imaginary values of γ. Putting $\beta^2-\alpha^2 = \frac{8\pi^2 m}{h^2}(U_a+U_b-2E)$ and $\beta_0^2-\alpha_0^2 = \frac{8\pi^2 m}{h^2}(U_a+U_b-2E_0)$ we thus obtain the following approximate expression for the energy difference $E-E_0$ as a function of γ,

$$E-E_0 = -\kappa\cos\gamma c, \tag{153}$$

with

$$\kappa = \frac{h^2(\alpha_0^2+\beta_0^2)}{8\pi^2 m}e^{-\beta_0 b}. \tag{153a}$$

For small values of the parameter γ corresponding to 'effective' wave-lengths $\lambda_{\text{eff}} = 2\pi/\gamma$ which are very large compared with c, $\cos\gamma c$ can be replaced by $1-\frac{1}{2}\gamma^2c^2$, so that (153) reduces to the form

$$E = \text{const.}+\tfrac{1}{2}\kappa c^2\gamma^2. \tag{154}$$

The second term is identical with the (kinetic) energy of a particle moving in a force-free space with a wave-length $\lambda = 2\pi/\gamma$, if the mass of this particle is defined by

$$m_{\text{eff}} = \frac{2m}{(\alpha_0^2+\beta_0^2)c^2}e^{\beta_0 b} \tag{154a}$$

and the velocity by

$$v = \frac{h\gamma}{4\pi m}(\alpha_0^2+\beta_0^2)c^2e^{-\beta_0 b}. \tag{154b}$$

so that

$$m_{\text{eff}}v = h\gamma/2\pi \quad \text{and} \quad m_{\text{eff}}v^2 = \kappa c^2\gamma^2.$$

This velocity is easily seen to coincide with the *group velocity*, which is obtained from the relation (153) with the help of the equation $v = d\nu/dk$. We have in fact, on replacing ν by E/h and k by $\gamma/2\pi$,

$$v = \frac{2\pi}{h}\frac{dE}{d\gamma} = \frac{2\pi\kappa c}{h}\sin\gamma c, \tag{155}$$

which for small values of γ reduces to (154b).

It should be mentioned that the group velocity as defined by (155) or (154b) is the wave-mechanical equivalent of the average velocity $v = c/s\tau$ with which the 'electron-wheel' is rolling according to the corpuscular picture sketched at the beginning of this section. The expression

$$\frac{h\gamma}{4\pi m}(\alpha_0^2+\beta_0^2)c^2e^{-\beta_0 b}$$

replaces the corpuscular expression $1/s\tau$ for the probability that the electron will pass in unit time from one atom to the next (s being the average or probable number of revolutions about the same atom).

The equations (153) and (155) were derived in 1928 by F. Bloch and by R. Peierls with the help of the perturbation theory for the case of a one-dimensional periodic field of force of any kind. The corresponding formulae for a three-dimensional lattice with mutually perpendicular axes are

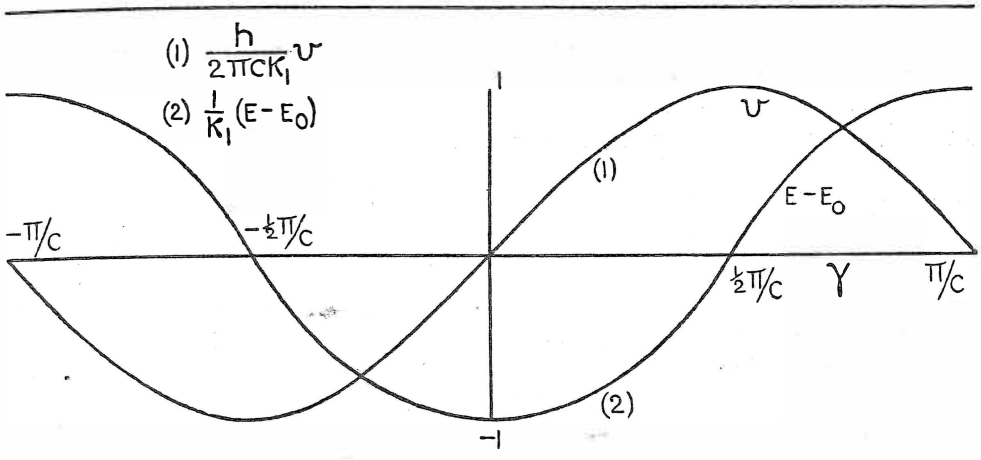

FIG. 18

$$E-E_0 = -\kappa_1 \cos\gamma_x c_1 - \kappa_2 \cos\gamma_y c_2 - \kappa_3 \cos\gamma_z c_3,$$

$$v_x = \frac{2\pi\kappa_1 c_1}{h}\sin\gamma_x c_1;\quad v_y = \frac{2\pi\kappa_2 c_2}{h}\sin\gamma_y c_2;\quad v_z = \frac{2\pi\kappa_3 c_3}{h}\sin\gamma_z c_3,$$

where c_1, c_2, c_3 are the spacings of the lattice (periodicity constants) for the respective axes.

The dependence of $E-E_0$ and v upon γ is shown graphically in the figure for the interval $-\pi/c \leqslant \gamma \leqslant \pi/c$, both functions being periodic with the same period $2\pi/c$ in γ. Actually, values of γ which differ from each other by an integral multiple of $2\pi/c$ are physically equivalent.

It is interesting to note that while for small values of γ the group velocity v increases with γ, that is, with the energy, for larger values lying between $\pi/2c$ and π/c, v is a decreasing function of E. In the neighbourhood of $\gamma = \pi/c$ we can put $\gamma = \pi/c - \gamma'$, where $\gamma' c$ is a small number, and replace $\cos\gamma c$ by $-\cos\gamma' c = -1+\frac{1}{2}\gamma'^2 c^2$. We thus get for E an expression of the form (154) with γ^2 replaced by $-\gamma'^2$, which is equivalent to the introduction of a negative effective mass.

We shall not go into further detail here and shall now come back to the question, hitherto left aside, as to the distribution of a large number

of electrons among the different stationary states of motion in the crystal lattice in accordance with the statistics of Pauli and Fermi. For this it will be necessary to consider a system consisting of a *limited* number of electrons in a crystal lattice of finite dimensions. So long as the self-consistent field can be assumed to be *the same for all the electrons*, this question is answered by the general formula:

$$\overline{N_k} = \frac{1}{e^{\mu(W_k - W_0)} + 1},$$

where W_k stands for the energy E of a quantized state and W_0 for a certain constant, determined by the condition $\sum_k \overline{N_k} = N$ and usually referred to as the 'chemical potential' of the electron gas. In the limiting case $T = 0$ ($\mu = \infty$) we find the same condition as for electrons moving in force-free space (bounded by impermeable walls), all the N states for which $W_k \leqslant W_0$ being occupied by one electron each and all the other states vacant. We shall illustrate this distribution law for the case of a one-dimensional model of a metal (or indeed any other solid body) consisting of a chain of n atoms. If this chain is open, the stationary states will be represented by standing waves of the type

$$\psi = C'(x)\cos\gamma x + C''(x)\sin\gamma x,$$

x being the distance reckoned along the chain, and C', C'' two periodic functions with the period c. Such waves can be obtained by the superposition of running waves of the type $C(x)e^{i\gamma x}$ with opposite values of γ, the magnitude of γ being fixed by the condition that the ends of the chain $x = 0$ and $x = nc$ should correspond to nodal points. This leads to $\psi = C(x)\sin\gamma x$ with

$$\gamma = \frac{2\pi r}{nc}, \tag{156}$$

where r is a whole number varying from zero to $n-1$. Values of r lying beyond this range do not actually lead to new states, since states for which γ differs by an integral multiple of $2\pi/c$ are physically undistinguishable.

Corresponding to each state of an electron in a separate atom (that is, for our model, corresponding to each energy-level of an electron enclosed in a valley of length a with infinitely high walls) we thus obtain a band formed by n closely lying states of motion in the lattice, whose energies can be expressed approximately by the equation (153) with the above n values of γ:

$$E_r = \text{const.} - \kappa\cos\frac{2\pi r}{n} \qquad (r = 0, 1, \ldots, n-1). \tag{156a}$$

If the number of electrons N is equal to that of atoms—as for alkali metals—and if each 'lattice-state', just as each 'atomic state', can only be occupied by two electrons with opposite spins, then all the $\frac{1}{2}n$ lower states of motion in the lattice (represented by the negative half of the dotted energy curve on the preceding fig.) will be occupied, while the other $\frac{1}{2}n$ states will be vacant. If there are two electrons per atom ($N = 2n$, alkaline-earth metals), all the n states of the band will be occupied. With more than two electrons per atom we must take into account not *one* state of the atom but at least two, each of these, if they lie sufficiently far apart, being replaced in the crystal lattice by a band consisting of n more or less closely lying 'lattice-states'; $2n$ electrons are accommodated in the first band (forming, as it were, the first floor of a big apartment house), and the others in the higher band or bands. It must be borne in mind, however, that this one-to-one correspondence between the 'atomic states' and the bands of 'lattice-states' can subsist only when the spacing between the atomic states (measured on the energy scale of course) remains large compared with the width of the lattice bands. If this condition is not fulfilled the different atomic states become completely amalgamated and interwoven in the resulting lattice states.

The preceding considerations which refer to the temperature of the absolute zero can be easily extended to $T > 0$. An interesting feature of this extension is that for narrow bands of lattice-states, with a number of electrons insufficient to fill one of them (for instance, one electron per atom), it may lead to a peculiar anomaly in the temperature dependence of the specific heat of the body, due to the switching of the electrons from the lower half of a given band to the higher half.† If the width of the band is of the order of magnitude of one or a few hundredth parts of a volt, this switch must be accompanied by a gradual increase of the specific heat with rise of the temperature over the value corresponding to the heat vibrations of the lattice followed by a decrease to the normal value after practically all the electrons have moved to the higher portion of the band. This is exactly what has been observed by Simon and others for many solid bodies.

32. The Calculation of the Self-Consistent Field in a Metal and on its Surface

In the preceding sections the 'self-consistent' field of force acting on an electron in a crystal lattice was supposed to be known. As a matter of

† This circumstance was pointed out to me by V. Sorokin.

fact all we know about it is its periodic character. A comparatively simple method for the approximate determination of the self-consistent field for a large number of electrons held together by some external field of force has been proposed independently by L. H. Thomas and E. Fermi for the special case of a heavy atom or ion. In this case the external field reduces to the attraction of the electrons to the positive nucleus. The method of Thomas and Fermi can, however, be easily extended to the case of an attractive external field of any kind.

It should be mentioned that the same method was used previously in connexion with the classical statistics for the determination of the distribution of electrons near the surface of an incandescent metallic body emitting them (Laue and others) or of the ions in the solution of an electrolyte about an electrically charged body (Stern, Debye).

Let the external field acting upon the electrons be specified by an electric potential $\phi_0(x,y,z)$. The mutual repulsion of the electrons in the equilibrium distribution can be replaced approximately by an adequately determined self-consistent external field with a potential $\phi_1(x,y,z)$. This potential or the total potential $\phi = \phi_0+\phi_1$ of the resulting equivalent field can be determined from Poisson's equation

$$\nabla^2\phi = +4\pi\epsilon n, \tag{157}$$

where $-\epsilon$ is the charge of an electron and n the average or probable number of electrons per unit volume. Now according to the classical statistics this number is in its turn determined as a function of ϕ by Boltzmann's formula

$$n = Ce^{\epsilon\phi/kT}, \tag{157 a}$$

where C denotes a constant. Inserting this in (157) we obtain an equation which can serve for the determination of ϕ as a function of the coordinates. This is the classical argument.

A quite similar scheme can be used in the Pauli-Fermi statistics which we must apply when dealing with an electron gas with a very high concentration n. The situation turns out, however, to be much more complicated if we are looking for a distribution corresponding to $T > 0$. In order to obtain n, i.e. the concentration of the electrons in the ordinary space, we must take the formula (141) of § 29, with $W = g^2/2m-\epsilon\phi$ and $d\gamma = 4\pi g^2\, dg dV$ and integrate over all values of g. Dividing the result by dV, we get

$$n = \frac{8\pi}{h^3}\int_0^\infty \frac{g^2\, dg}{e^{\lambda+\mu(g^2/2m-\epsilon\phi)}+1}.$$

The integration is very easily carried out, leading to the formula (157 a) if 1 is dropped in the denominator of the integrand—which corresponds to the classical theory. It cannot be carried out, however, in finite terms so long as the first term in the denominator is small compared with 1—which is the case for not too large values of g for an electron gas of high concentration.

The latter condition, which is characteristic of the distribution of the electrons both in a heavy atom and in a solid body, leads, however, to a great simplification of the theory, for it permits us, so long as we are dealing with ordinary temperatures, to replace this temperature by zero. Writing $\lambda = -\mu W_0$, where W_0 is a constant, the 'chemical potential' of the electron gas, the denominator of the integrand in (158) vanishes for $g^2/2m - \epsilon\phi > W_0$ and reduces to unity for $g^2/2m - \epsilon\phi < W_0$. We then get

$$n = \frac{8\pi(2m)^{\frac{3}{2}}}{3h^3}(\epsilon\phi + W_0)^{\frac{3}{2}}. \tag{157 b}$$

Substituting this in (157) we obtain the fundamental equation of the Thomas-Fermi theory

$$\nabla^2 f = \gamma f^{\frac{3}{2}}, \tag{158}$$

where

$$f = \epsilon\phi + W_0 \quad \text{and} \quad \gamma = \frac{32\pi^2}{3}\epsilon^2\left(\frac{2m}{h^2}\right)^{\frac{3}{2}}. \tag{158 a}$$

The equation (158) must be integrated under certain boundary conditions, which for an infinite lattice can be replaced by the condition of periodicity, together with the conditions specifying the external field ϕ_0 and the total number of electrons N. The latter serves also for the determination of W_0. It should be mentioned that the function $\epsilon\phi + W_0$ must have non-negative values only. If it vanishes on a certain closed surface it must vanish also outside it.

A solution of equation (158) for a three-dimensional crystal lattice has not yet been obtained.† It can easily be obtained, however, if we simplify our problem, replacing the positive nuclei or rather ions by a *uniform distribution of positive electricity* extending to the boundary of the body under consideration. This simplification can be allowed for metals if the (valence) electrons are dealt with as completely free within the body, where the potential must then have a constant value, C say. This means that within the body the electrons will form a uniform

† For a two-dimensional lattice the corresponding problem can be easily solved, thanks to the fact that the Fermi distribution law when applied to a motion in *two* dimensions leads to a *linear* equation $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \gamma' f$ instead of (158). Cf. Lennard Jones, *Proc. Roy. Soc.* **120**, 727 (1928).

distribution of negative electricity with a density of the same magnitude ρ as that of the positive one. Assuming one (valence) electron per atom, we have

$$\rho = \epsilon n.$$

This relation will, however, be altered at and near the surface of the body, where the electrons will tend to escape, protruding beyond the surface S, which limits the uniform distribution of the positive charge (and which can be defined as the surface of the body). They will thus form a layer of negative charge outside this surface, compensated by a layer of positive charge on the inner side of the latter, where there will be a shortage of negative electricity. The metallic body will thus appear coated by an electric double layer (or 'condenser') of very small thickness with a positive inner side and a negative outer one, the potential C inside it being equal to the potential jump across this layer if the potential of the outer space (at infinity) is assumed to be zero.

The actual distribution of potential and charge in this surface double layer of our simplified model can be easily calculated by means of the equation (158) if the surface of the body is dealt with as an infinite plane, $x = 0$ say. If the body extends in the negative direction we must have $\phi = C$ for $x = -\infty$ and $\phi = 0$ for $x = \infty$. The latter condition in connexion with the relation (157 b) and the obvious condition that $n = 0$ at $x = \infty$ shows that in our case (of a neutral metallic body) the 'chemical potential' W_0 is equal to zero.

The equation (158) is thus to be integrated under the conditions $f = 0$ at $x = \infty$ and $f = C\epsilon$ at $x = -\infty$. Since f is a function of x only, it reduces to

$$\frac{d^2f}{dx^2} = \gamma f^{\frac{3}{2}}. \tag{159}$$

Multiplying it by df/dx, we get

$$\frac{d}{dx}\frac{1}{2}\left(\frac{df}{dx}\right)^2 = \tfrac{2}{5}\gamma\frac{d}{dx}f^{\frac{5}{2}},$$

that is,

$$\left(\frac{df}{dx}\right)^2 = \tfrac{4}{5}\gamma f^{\frac{5}{2}} + \text{const.}$$

For the positive side ($x > 0$) the constant of integration must be set equal to zero (since df/dx vanishes along with f at $x = \infty$). In extracting the square root we must further take the negative sign, for f is obviously decreasing with increase of x. This leads to

$$f^{-\frac{1}{4}} - f_0^{-\frac{1}{4}} = \tfrac{1}{2}(\gamma/5)^{\frac{1}{2}}x,$$

or
$$f(x) = \frac{1}{\{f_0^{-\frac{1}{2}} + \frac{1}{2}(\gamma/5)^{\frac{1}{2}}x\}^4} \qquad (x > 0) \tag{159a}$$

where f_0 denotes the value of f at $x = 0$.

For the negative side ($x < 0$) we get, similarly,

$$\left(\frac{df}{dx}\right)^2 = \tfrac{4}{5}\gamma(f^{\frac{5}{2}} - f_{-\infty}^{\frac{5}{2}})$$

and
$$\int_{f_0}^{f} \frac{df}{(f_{-\infty}^{\frac{5}{2}} - f^{\frac{5}{2}})^{\frac{1}{2}}} = \sqrt{(\tfrac{4}{5}\gamma)}\,|x| \qquad (x < 0), \tag{159b}$$

the integration leading to a transcendental function which need not be considered here. $f_{-\infty} = C\epsilon$ denotes the value of f at $x = -\infty$. In order to secure the continuity of df/dx at $x = 0$ we must put

$$f_0 = \tfrac{1}{2}f_{-\infty} = \tfrac{1}{2}C\epsilon,$$

as of course might be expected without any calculation.

The magnitude of C can be obtained from (157 b) if $n_{-\infty}$ is known, being equal to the volume concentration of the valence electrons in the metal sufficiently far away from its surface. Using the value $n_{-\infty} = 2.10^{22}$ for sodium we get $\phi_{-\infty} = C \cong 3$ volts. The effective thickness of the corresponding double layer, as measured by the expression $\frac{4f_0^{-\frac{1}{4}}}{\sqrt{(\gamma/5)}}$ (according to (159a)), turns out to be of the order of magnitude of 6.10^{-9} cm., that is, of the same order of magnitude as the distance between the neighbouring atoms in a solid body.

The value of C obtained by the preceding method is of course too low, since we have seen before that the work of escape for an electron initially at rest, which should be measured by the product $C\epsilon$, must lie between 10 and 20 volts. It should further be pointed out that according to the derivation of the equation (157 b) $\epsilon\phi + W_0$ must be considered as the maximum kinetic energy of an electron at a given place. With $W_0 = 0$ this maximum kinetic energy is equal to $\epsilon\phi$, which means that the fastest electrons have just enough energy to enable them to escape from the metal even at the temperature of the absolute zero. This obviously erroneous result is due to the failure of the theory to determine the actual field of force acting on an electron, the actual field being replaced by a smoothed field in the creation of which the electron under consideration is also supposed to partake, with its charge diluted over a relatively large volume.

Similar results are found in applying the theory of Thomas and Fermi to the distribution of electrons (or potential) in a heavy atom,

the equation (159) being replaced in this case, in view of the spherical symmetry of the problem, by

$$\frac{d^2(rf)}{dr^2} = \gamma r f^{\frac{3}{2}}, \tag{160}$$

where r is the distance from the nucleus. With $r\phi = 0$ at infinity—which corresponds to a neutral atom—we find again $W_0 = 0$ and the work of escape—or ionization potential—is equal to zero. This discrepancy does not play any role, however, so long as we are interested in the distribution of electricity in the deeper regions of the atoms—the Thomas-Fermi theory leading in fact to excellent agreement with experimental data derived from X-ray scattering. But it shows that this theory can hardly be applied to the periphery of an atom, and accordingly, hardly to the regions of a material body lying in between the neighbouring atoms.

The 'intrinsic' electric potential of a metal, and indeed of any other material body, with respect to the external space can be computed in a much simpler way on the assumption that the electron distribution about each atom remains unaltered by their lying close together. If the surface of the body is defined as the surface passing through the nuclei of the outermost atoms, there will be a layer of negative charge on the outer side of it due to the electrons which are moving about these outermost nuclei and a compensating positive charge on the inner side. It can easily be shown that the potential jump across such an electric double layer is equal to

$$\phi = 2\pi n\epsilon \sum \overline{r^2}, \tag{161}$$

where r is the distance of an electron from the corresponding nucleus, and $\sum \overline{r^2}$ the sum of the average values of $\overline{r^2}$ for all the electrons in an atom.†

The same formula can be obtained by averaging the electric potential within the body on the assumption that the distribution of negative electricity about each nucleus remains spherically symmetrical. Let us imagine that an atom consists of a nucleus with a charge $Z\epsilon$ and a compensating spherical layer of negative electricity with a radius r. The electric potential within this atom (which can be pictured as a minute spherical condenser) is equal to $Z\epsilon/R$ $(R < r)$, while outside it is equal to zero. If the average volume occupied by an atom in the body under consideration is v_1, then the average value of the potential

† Cf. J. Frenkel, *Phil. Mag.*, April 1917.

in this body can be defined as

$$\frac{Z\epsilon}{v_1}\int_0^r \frac{4\pi R^2\,dR}{R} = \frac{2\pi Z\epsilon r^2}{v_1},$$

or $2\pi n Z\epsilon r^2$, since v_1 is equal to the reciprocal of the number n of atoms in unit volume. Replacing Zr^2 by the sum $\sum \overline{r^2}$ for all the electrons, we obtain formula (161). According to the wave-mechanical theory we must put

$$\overline{r^2} = \int r^2 \psi\psi^*\,dV, \tag{161 a}$$

where ψ denotes the wave function of the corresponding electron.

The above theory does not take into account that alteration of the electron distribution which is due to the mutual action of the atoms and which must be particularly important on the surface of the body. A more effective method of solving the problem under consideration is to start with that distribution of negative electricity which is obtained wave-mechanically on the simple assumption that the potential energy of an electron is constant within the metal and suddenly increases by a certain amount U on passing through its surface from the inside to the outside. So long as the total energy of the electrons E_1 is smaller than U, this leads to wave functions which in the external space ($x > 0$) have the form $\psi = Be^{-\beta x}$ with $\beta = \frac{2\pi}{h}\sqrt{\{2m(U-E_1)\}}$. This corresponds to a probable volume density of electric charge

$$\rho = -\epsilon\psi\psi^* = -\epsilon|B|^2 e^{-2\beta x} \tag{162}$$

for each separate electron. We thus see that at least a certain part of the layer of negative electricity extending beyond the surface of the metal (i.e. the discontinuity surface of the potential energy) can be traced wave-mechanically to the total reflection of the electron waves impinging on the surface from the inside. In order to find the negative charge due to all the electrons penetrating in this way through the surface we must add the individual electron functions $|\psi|^2$, taking into account the Fermi distribution law for the velocities (normal to the surface). We shall not linger over the mathematical development of this scheme, and shall limit ourselves to showing how it can be used for the actual calculation of U. Let us assume, for the sake of simplicity, that the resulting distribution of charge is given by

$$\rho = \rho_0 e^{-2\bar{\beta}x} = -\epsilon n e^{-2\beta x}, \tag{162 a}$$

where β is some average value of β containing U as a parameter.

Introducing a compensating positive layer of uniform density on the inner side of the surface, we can determine the drop of potential ϕ through the resulting double layer with the help of Poisson's equation $\frac{d^2\phi}{dx^2} = -4\pi\rho$. It seems now natural to identify the corresponding change of potential energy $\epsilon\phi$ with the value U which forms the starting-point of our calculation (in agreement with the idea of the self-consistent field). This leads to an equation for the actual determinations of U. The values of U so obtained, though larger than those obtained by the Thomas-Fermi method, are nevertheless too small. In fact, as shown by Ig. Tamm, they are not sufficient to prevent the fastest electrons from escaping outside.

It should be pointed out, however, that the identification of U with $\epsilon\phi$ is actually not allowed for the following reason: In the theory of the self-consistent field the potential ϕ refers, strictly speaking, not to one of the electrons which contribute to it but to an imaginary electron which penetrates into the body without disturbing the distribution of charge and potential produced by the real electrons.

This difference can be described by adding to $-\epsilon\phi$ (the potential energy of the imaginary electron inside the metal) a potential energy $U_0' = -\epsilon^2/4x_0$ corresponding to the work of the 'image force', where x_0 is a certain characteristic distance of the order of magnitude of the thickness of the unperturbed surface double layer. We thus get instead of $U = \epsilon\phi$

$$U = \epsilon\phi + \frac{\epsilon^2}{4x_0}.$$

The notion of the image force as applied to an electron is, however, open to objection. In fact, this force is due to a modification of the distribution of electrons near the surface when one of the electrons is supposed to be held at a definite point near the surface. Now, as a matter of fact, the electron is not held in a fixed position but is moving with a relatively very large velocity either to the outside or to the inside. Under this condition the distribution of electrons inside the metal (near the surface) will not be altered 'adiabatically', that is, in such a way as if the electron in question remained at a fixed point. It will not be possible accordingly to define the (additional) force acting on the electron at a given distance x from the surface irrespective of its velocity, the 'image force' thus becoming an 'imaginary force', as it were.—This criticism does not apply to the case of a proton or an ion with a still larger mass, moving with a kinetic energy of the same

order of magnitude as that of the free electrons in the metal, the velocity of the proton or ion being then much smaller than that of the electron.

33. Electric Properties of Contacts between Conducting Bodies

Although unable to escape from a metal under ordinary conditions (on account of $U > W$),† the 'free' electrons can be extracted from it by an intense electric field pulling them outside as has been shown in § 16, or by light of sufficiently high frequency determined by the condition $h\nu \geqslant U - W_{\max}$. They can also escape spontaneously if the temperature of the metal is high enough so that a measurable fraction of all the electrons while moving inside the metal possess a kinetic energy larger than the work of escaping, U. In order that an electron might escape through the surface $x = 0$, the part of its kinetic energy corresponding to the x-component of the velocity must be larger than U, while the value of the two other components of the velocity parallel to the plane $x = 0$ is immaterial.

Denoting the number of electrons per unit volume whose momentum components lie within assigned limits by $f\, dg_x\, dg_y\, dg_z$ we thus find for the total number of electrons escaping through unit area of the plane $x = 0$, in unit time, the expression

$$I = \int_{g_x^0}^{\infty} \frac{g_x}{m}\, dg_x \int\int_{-\infty}^{+\infty} f\, dg_y\, dg_z, \tag{163}$$

where g_x^0 is defined by

$$\frac{1}{2m}(g_x^0)^2 = U, \tag{163 a}$$

while the distribution-function f is given, according to (141), by

$$f = \frac{2}{h^3} \frac{1}{e^{\mu(W - W_0)} + 1} \qquad \left(\mu = \frac{1}{kT}\right) \tag{163 b}$$

with $\lambda = -\mu W_0$. As has been shown above, for not too high temperatures, up to a few thousand degrees, W_0 preserves approximately the same value as for $T = 0$, i.e. remains equal to the maximum kinetic energy of the electrons (inside the metal) at the temperature of absolute zero.

Putting $W = W_x + W_r$, where $W_x = g_x^2/2m$ denotes that part of the kinetic energy which corresponds to the component of motion normal to the surface, while $W_r = (g_y^2 + g_z^2)/2m$ denotes the part corresponding to the motion parallel to it, and introducing in the momentum space

† W denotes here, as before, not the total but the *kinetic* energy of an electron.

a cylindrical coordinate system with its axis in the g_x direction, we get

$$\int_{-\infty}^{+\infty}\int f\, dg_y\, dg_z = 2\pi m \int_0^\infty f\, dW_r,$$

and

$$I = 2\pi m \int_U^\infty dW_x \int_0^\infty f\, dW_r. \tag{164}$$

Now the integral

$$\int_0^\infty f\, dW_r = \frac{2}{h^3}\int_0^\infty \frac{dW_r}{e^{\mu(W_r+W_x-W_0)}+1}$$

is easily evaluated with the help of the substitution

$$e^{\mu(W_r+W_x-W_0)} = \xi, \qquad dW_r = \frac{1}{\mu}\frac{d\xi}{\xi},$$

so that

$$\int_0^\infty f\, dW_r = \frac{2}{h^3\mu}\int_{\xi_0}^\infty \frac{d\xi}{\xi(\xi+1)} = \frac{2}{h^3\mu}\log\frac{\xi_0+1}{\xi_0} = \frac{2}{h^3\mu}\log(1+\xi_0^{-1}), \tag{164a}$$

where

$$\xi_0 = e^{\mu(W_x-W_0)}. \tag{164b}$$

For values of W_x which are larger than U this expression must be exceedingly large (the difference $U-W_0$ being of the order of magnitude of a few volts, while $kT = 1/\mu$, even for temperatures of a few thousand degrees, amounts but to a few tenths of a volt). We can accordingly, in carrying out the integration with respect to W_x in (164), put

$$\log(1+\xi_0^{-1}) = \xi_0^{-1}.$$

We thus get

$$I = \frac{4\pi m}{h^3\mu}e^{\mu W_0}\int_U^\infty e^{-\mu W_x}\, dW_x,$$

that is

$$I = \frac{4\pi m}{h^3}(kT)^2 e^{-\frac{U-W_0}{kT}}. \tag{165}$$

This expression multiplied by the electronic charge ϵ gives the (maximum) value of the 'thermionic' current from an incandescent metal. This current was first discovered by Edison in connexion with his experiments on the electric lamp and thoroughly studied by Richardson, who derived for it, from thermodynamical considerations, the expression

$$I = AT^2 e^{-W/kT} \qquad (A = \text{const.}),$$

which is equivalent to (165). The latter formula has been obtained by Sommerfeld. The theoretical value of Richardson's constant

$$A = \frac{4\pi m k^2 \epsilon}{h^3}$$

corresponds to 120 amperes/cm.^{2}, and is nearly twice as large as the experimental value. This discrepancy, as pointed out by Fowler and Nordheim, can be accounted for by the fact that on the wave-mechanical theory the electrons have a certain probability ($\simeq \frac{1}{2}$) of being reflected from the surface of the metal even if they strike it with a kinetic energy which is sufficient for transmission.

Our initial expression (164) for I should be modified accordingly by multiplying the distribution function f with a transmission coefficient D which can be defined as a certain function of W_x. We thus get, instead of (164),

$$I = 2\pi m \int_0^\infty D(W_x)\, dW_x \int_0^\infty f(W_x + W_r)\, dW_r. \tag{166}$$

We shall not stop to give a more exact evaluation of I for the case just considered, and shall now turn to the consideration of another interesting case, that of two metallic bodies in 'contact' with each other.

According to the usual conception such a contact, inasmuch as it is characterized by the fact that the electrons are able to pass from one body to the other, is visualized as a junction where the distance apart between the two bodies is of the same order of magnitude as the distance between the nearest atoms in each of them. Now according to the wave-mechanical theory the electrons will be able to pass from one metal to the other even when they are separated by a *gap* much larger than the interatomic distance, and this is true at *any* temperature, even that of the absolute zero. The waves representing them will suffer a total reflection at the surface of the corresponding metal but nevertheless penetrate with an exponentially decreasing amplitude into the second metal, where the motion of the electrons will resume its normal character.

As a matter of fact, when two metallic surfaces are brought into contact with each other, actual contact can exist as a rule only for a very small portion of these surfaces, whereas the major part of them may be separated by gaps many atomic distances wide. From the point of view of the conduction of electricity, these gaps must also be taken into account—unless they are too wide.

In order to investigate the dependence of the electrical resistance of a contact-gap upon its width, we shall consider an idealized case of two metals a and b with absolutely plane surfaces separated from each other by a gap of constant width δ. The first metal will thus be imagined to extend from $x = -\infty$ to $x = 0$, and the second from $x = \delta$ to $x = \infty$.

Under these conditions there will be an exchange of electrons between the two metals, the number of electrons flowing in each direction per unit time and unit area being represented by (167) where the transmission coefficient D must be suitably defined as a function of W_x.

According to the general law of reversibility, the coefficient D must have the same value for a transition from a to b, with an initial kinetic energy W_{xa} and a final energy W_{xb}, as for the inverse transition from b to a with an initial kinetic energy W_{xb} and a final kinetic energy W_{xa}. We thus have

$$D_{ba}(W_{xa}+U_{ab}) = D_{ab}(W_{xa}),$$

where U_{ab} denotes the difference $W_{xb}-W_{xa}$, i.e. the decrease of potential energy of an electron passing from a to b.

Using the expression

$$I_{ab} = 2\pi m \int_0^\infty D_{ab}(W_x)\, dW_x \int_0^\infty f_a(W_x+W_r)\, dW_r$$

for the number of electrons passing from a to b, we get accordingly for the number of electrons passing from b to a the expression

$$I_{ba} = 2\pi m \int_0^\infty D_{ab}(W_x)\, dW_x \int_0^\infty f_b(W_x+U_{ab}+W_r)\, dW_r,$$

so that the resulting flow of electrons in the direction $a \to b$ per unit time and area is given by

$$I = \frac{2\pi m}{h^3} \int_0^\infty D_{ab}(W_x)\, dW_x \int_0^\infty [f_a(W_x+W_r)-f_b(W_x+W_r+U_{ab})]\, dW_r. \quad (167)$$

This formula shows incidentally that I vanishes, that is, that the two metals are in equilibrium if, and only if,

$$f_a(W_a) = f_b(W_a+U_{ab}) = f_b(W_b). \quad (168)$$

Taking into account the expression (164 b) and replacing W_0 in the distribution function by W_{0a} and W_{0b} respectively, we see that this equilibrium condition can actually be fulfilled so long as the two metals have the same temperature, if

$$U_{ab} = W_{0b}-W_{0a}. \quad (168\,a)$$

This relation expresses the equality of the chemical potentials or free energies of an electron in the two metals and indeed could be obtained by purely thermodynamical considerations, without the use of the reversibility condition for D_{ab} and D_{ba}.

If U_a and U_b denote the potential energy jumps through the surface

of each metal, and V_{ab} the *decrease* of potential energy (in the direction $a \to b$) due to the presence of an electric field in the gap between the two bodies, then

$$U_{ab} = U_b - U_a - V_{ab}, \tag{169}$$

as is easily seen from Fig. 19, where the line M, N, Q, R, S, T represents the potential energy as a function of x (the line NQ corresponding to the plane $x = 0$ and the line RS to $x = \delta$). The equilibrium condition (168 a) can thus be rewritten in the form

$$V_{ab} = (U_a - W_{0a}) - (U_b - W_{0b}). \tag{169 a}$$

The 'external' potential energy difference V_{ab} is equal to the product of the electronic charge ϵ by the potential difference between the two metals. The equilibrium value of this potential difference is what is called the *contact potential difference*. According to the preceding equation and Sommerfeld's formula (165) it is equal to the difference of the 'work functions' $w = U - W_0$ which specify the electronic emission from each metal at high temperatures. This result is in full agreement with the experimental evidence.

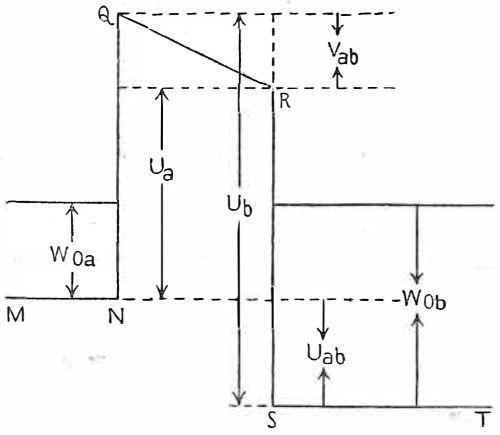

FIG. 19

The contact potential difference between two different metals is established automatically by the mere fact that the resulting flow of electrons between them in the absence of external forces stops only when the equilibrium value corresponding to (169 a) is reached.

The preceding results can be visualized by regarding the electrons as a fluid distributed on the preceding energy diagram (Fig. 19) with a variable density (in the W_x space), which in the equilibrium condition has the same value for the same horizontal level on both sides of the gap between the two metals, so long as this level lies above the two bottom lines (MN and ST). At the absolute zero this density can be regarded as constant from top to bottom (the total number of electrons per unit volume being in the ratio $W_{0a}^{\frac{3}{2}} : W_{0b}^{\frac{3}{2}}$). For higher temperatures it is variable, coinciding, however, for both metals on the side of the larger velocities, where the same 'Maxwell tail' is obtained. This condition can be also pictured by the full line in the next figure where OAA_1B represents the Fermi curve for the metal b and O_1A_1B for the

metal a, the distance OO_1 being equal to U_{ab}. The portions of the two curves between O_1 and infinity exactly coincide.

If an additional electric field E is applied across the gap increasing the equilibrium value of V_{ab} by $V' = \epsilon E\delta$, one of the two curves is shifted with respect to the other along the energy axis by the amount V', as shown, for example, by the dotted line $O'A'B'$. This will result in a current I flowing from b to a (or from a to b) due to the electrons represented by the area between A_1B and A_1B'. For small values of V' this current will be obviously proportional to V' and will have practically the same magnitude for opposite values of V'.

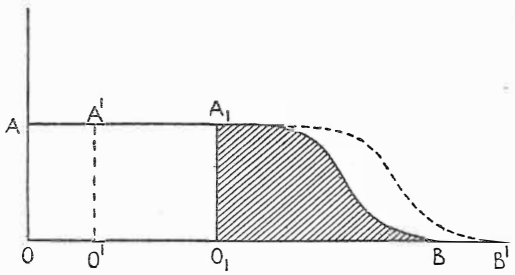

FIG. 20

With larger values of V' the function $I(V')$ will increase more rapidly, remaining practically symmetrical with regard to the sign of V' until values of V' are reached which are comparable with the smaller of the two values W_{0a} and W_{0b}, i.e. with W_{0x} in our case.

These properties of the function $I(V')$, which are characteristic of contacts, serve to explain a number of various phenomena connected with the passage of electricity through contacts between conducting bodies—such as the relatively high resistance of thin metallic films and its anomalous variation with the temperature, the deviations from Ohm's law in the conduction of electricity through compressed powders of poorly conducting substances, and finally for the *rectification phenomena*, which are observed with contacts between a metal and semiconductor.

In the case of two metals such rectification (i.e. asymmetrical change of $I(V')$ with the change of sign of V') has not been observed for the probable reason that the 'chemical potentials' W_0 of different metals are all of the same order of magnitude (about 10 volts) and very large compared with the potential differences V' applied across the contact gaps. A semi-conductor (such as Cu_2O, for example) can be considered, roughly speaking, as a metal with a relatively very small electron concentration (which varies with the temperature according to the law $e^{-Q/kT}$ characteristic of dissociative equilibrium). It must differ accordingly from a metal by a relatively very small value of the chemical potential W_0 (in fact the latter can become negative), a circumstance

which explains the possibility of a rectification for applied potential differences V' of a relatively small magnitude.

These results can be easily put into a quantitative form with the help of the expression (167). In the case of sufficiently small values of V' we can take avail of the relation (168) (which holds as long as the two values W_a and W_b remain positive) and put accordingly

$$f_b(W_b) = f_a(W_a+V'), \tag{170}$$

where V' is the excess of U_{ab} or V_{ab} over the equilibrium value (168 a). The expression (167) now assumes the form

$$I = 2\pi m \int_0^\infty D_{ab}(W_x)\, dW_x \int_0^\infty [f_a(W_x+W_r)-f_a(W_x+W_r+V')]\, dW_r. \tag{170 a}$$

For very small values of V' we can put with sufficient approximation

$$f_b(W)-f_b(W+V') = -V' \frac{\partial f_b}{\partial W_r},$$

which leads to

$$I = CV', \tag{171}$$

i.e. to Ohm's law, the conductivity of the contact C being equal to

$$C = 2\pi m \int_0^\infty D_{ab}(W_x) f_Q(W_x)\, dW_x. \tag{171 a}$$

The transmission coefficient $D(W_x)$ in this formula can be considered as a constant quantity corresponding to $V' = 0$. Its exact dependence on W_x is rather complicated so long as the line QR on the energy diagram Fig. 19 is inclined (which is the case with two different metals). Replacing it by a horizontal line passing through its centre, at a level

$$\overline{U} = U_a - W_{0a} + \tfrac{1}{2}V_{ab} = \tfrac{1}{2}[(U_a - W_{0a}) + (U_b - W_{0b})]$$

over the line $W_x = W_{0a}$ corresponding to the fastest electrons in a, and, putting $W_x - W_{0a} = \xi$, we get approximately

$$D_{ab}(W_x) = e^{-\gamma\delta\sqrt{(\overline{U}-\xi)}},$$

where $\gamma = \frac{4\pi}{h}\sqrt{(2m)}$. (Cf. § 16.)

It is clear from (171 a) that the conductivity C is raised but little with the temperature, so that the function $f_a(W_x)$ can be replaced by its limiting form corresponding to $T = 0$, i.e. $f_a = 2/h^3$ for $W_x < W_{0a}$ ($\xi < 0$) and $f_a = 0$ for $W_x > W_{0a}$ ($\xi > 0$). We thus have with $\xi = \xi'$

$$C \simeq \frac{4\pi m}{h^3} \int_0^{W_{0a}} e^{-\gamma\delta\sqrt{(\overline{U}+\xi')}}\, d\xi'. \tag{171 b}$$

It should be mentioned that so long as we limit ourselves to the case $T = 0$ the difference $f_a(W_x+W_r)-f_a(W_x+W_r+V')$ must either vanish or reduce to $2/h^3$ if $W_x+W_r < W_{0a}$ while $W_x+W_r+V'_\infty > W_{0a}$. We thus get in this case *exactly*

$$\int_0^\infty [f_a(W_x+W_r)-f_a(W_x+W_r+V')]\,dW_r = \frac{2}{h^3}V',$$

so that equation (171 a) remains valid for larger values of V', provided we take into account the dependence of D upon V.

Before we consider this question let us come back to the expressions (170) and (170 a). The latter corresponds to a positive current, that is, to a flow of electrons in the direction $a \to b$ if V' is positive, and to a negative current, i.e. to a flow in the reverse direction if V' is negative. Now in the simplest case of two identical metals the function $I(V')$ does not seem at first sight to be odd, since the expression $f(W)-f(W+V')$ does not alter its sign along with V'. This apparent lack of symmetry (or rather 'antisymmetry') is explained by the fact that with a negative value of V' the transmission coefficient $D_{ab}(W_x)$ is zero for values of W_x which are smaller than $|V'|$, whereas with $V' > 0$ it is different from zero for all values of W_x. If in the former case $D_{ab}(W_{xa})$ is replaced by $D_{ba}(W_{xb})$, where $W_{xb} = W_{xa}+V' = W_{xa}-|V'|$, and $f_a(W_a)$ by $f_b(W_b-V') = f_b(W_b+|V'|)$, we obtain instead of (170 a)

$$I = 2\pi m \int_0^\infty D_{ba}(W_x) \int_0^\infty [f_b(W_x+W_r+|V'|)-f_b(W_x+W_r)]\,dW_r,$$

which is equal and opposite to (170 a) if $|V'|$ is the same in both cases.

With a positive value of V' we can put

$$D_{ab}(W_x) \cong e^{-\gamma\delta\sqrt{(\overline{U}-\frac{1}{2}|V'|-\xi)}}, \tag{172}$$

the same formula being valid for D_{ba} with negative values of V'. So long as $\frac{1}{2}|V'|+\xi$ is small compared with $\overline{U}$ we can put

$$(\overline{U}-\tfrac{1}{2}|V'|-\xi)^{-\frac{1}{2}} = \sqrt{\overline{U}} - \frac{\frac{1}{2}|V'|+\xi}{2\sqrt{\overline{U}}},$$

and consequently

$$D_{ab}(W_x) \cong D_0\, e^{+\beta(\frac{1}{2}|V'|+\xi)}, \tag{172 a}$$

where

$$D_0 = e^{-\gamma\delta\sqrt{\overline{U}}} \quad \text{and} \quad \beta = \frac{\gamma\delta}{2\sqrt{\overline{U}}}. \tag{172 b}$$

This leads to the following approximate expression for the conductivity of the contact gap at $T = 0$:

$$C = C_0\, e^{\frac{1}{2}\beta|V'|}, \tag{173}$$

where C_0—the value of C for very small values of $|V'|$—is given by

$$C_0 = \frac{4\pi m}{h^3} D_0 \frac{e^{\beta W_0} - 1}{\beta}. \tag{173a}$$

The dependence of C upon $|V'|$ described by equation (173) seems to be in good agreement with the results of recent experiments on the conductivity of compressed powders of carborundum and some other substances (serving as high-tension dischargers). The experimental values of β correspond to gap-widths of about 10^{-7} cm.

Although the preceding expressions refer to two identical metals,

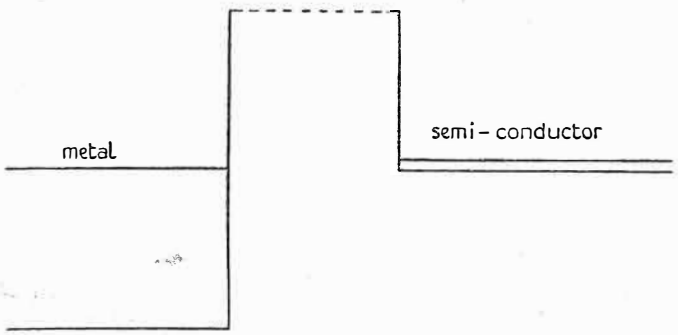

FIG. 21

they can be applied to contacts between different metals, so long as $|V'|$ is not too large.

In the case, however, of a contact between a metal and a semi-conductor they become inapplicable for relatively small values of $|V'|$. We must return in this case to the general equation (167) with $U_{ab} = U^0_{ab} + V'$ where U^0_{ab} is the equilibrium value of U_{ab} given by (169).

Since the electron concentration and consequently the 'chemical potential' W_0 is in the case of a semi-conductor very small compared with that of a metal, whereas the work function $U - W_0$ is in both cases of the same order of magnitude, the energy U^0_{ab} must have a rather large positive value if a denotes the semi-conductor and b the metal (so that in the equilibrium condition all the a-electrons will be able to pass into b, while but a relatively very small fraction of the b-electrons will be able to pass into a; see Fig. 21).

For the semi-conductor we shall have with sufficient approximation

$$f_a(W) = Ae^{-\mu W} = \frac{2}{h^3} e^{-\mu(W - W_{0a})}$$

(Maxwell's law being the limiting form of that of Fermi for a small

electron concentration), and consequently

$$\int_0^\infty f_a(W_x+W_r)\,dW_r = \frac{2}{h^3\mu}\,e^{-\mu(W_x-W_{0a})}.$$

The corresponding expression for the metal is given according to (164 a) and (164 b) by

$$\int_0^\infty f_b(W_x+W_r+U^0_{ab}+V')\,dW_r = \frac{2}{h^3\mu}\log[1+e^{-\mu(W_x+U^0_{ab}+V'-W_{0b})}],$$

or, since

$$U^0_{ab}-W_{0b} = -W_{0a}$$

and

$$W_x-W_{0a}+V' \gg kT,$$

$$\int_0^\infty f_b(W_x+W_r+U^0_{ab}+V')\,dW_r \cong \frac{2}{h^3\mu}\,e^{-\mu(W_x-W_{0a}+V')}.$$

We thus get

$$I = \frac{4\pi m}{h^3\mu}\,e^{\mu W_{0a}}\int_0^\infty dW_x\,D_{ab}(W_x)e^{-\mu W_x}(1-e^{-\mu V'}).$$

If the dependence of D_{ab} upon V' is taken into account, then

$$I = I_0\,e^{\frac{1}{2}\beta|V'|}(1-e^{-\mu V'}), \tag{174}$$

where I_0 is a function of the temperature alone of the same type as the number of free electrons in the semi-conductor, i.e.

$$I_0 = \text{const.}\;e^{-Q/kT}, \tag{174 a}$$

which we need not consider in detail.

The preceding formula is valid for positive values of V' only, driving the electrons from the semi-conductor into the metal. But for the factor $e^{\frac{1}{2}\beta|V'|}$, which corresponds to an increase of the transmission coefficient D_{ab} with the applied field, it would correspond to a saturation effect for sufficiently large values of V', due to practically all the free electrons of the semi-conductor being driven into the metal with no compensating stream in the reverse direction. It should be mentioned that this saturation current, so far as its dependence on the temperature is concerned, is very similar to the Richardson effect, i.e. to the electron emission from an incandescent metal into a vacuum, the work function of the metal w corresponding to the energy Q which determines the number of free electrons in the semi-conductor. (Q is approximately equal to one-half of the energy necessary for the dissociation of one such electron from the atom to which it is bound at low temperatures.) As a matter of fact the saturation current we are considering is an exact analogy of the Richardson effect, the free

electrons streaming from the metal into the vacuum being replaced by the bound electrons streaming from the semi-conductor into the metal through the contact gap, which they can pass with a deficient kinetic energy.

In order to obtain the current due to an electric field which is driving the electrons in the reverse direction, i.e. from the metal into the semi-conductor, we have but to reverse the sign of V' and of I in the expression (174). We thus get

$$I = I_0 e^{\frac{1}{2}\beta|V'|}(e^{\mu V'} - 1), \tag{174 b}$$

which corresponds to a practically exponential increase of I with V'.

The ratio

$$\frac{e^{\mu|V'|} - 1}{1 - e^{-\mu|V'|}}$$

could be defined as the 'theoretical rectification coefficient' of the contact under consideration. It should be borne in mind, however, that in practice this coefficient, i.e. the ratio of the magnitudes of I for equal and opposite values of the potential difference, refers not to the potential difference V' applied to the *contact*, but to the *total* potential difference V acting in the circuit which contains the contact.

To equal and opposite values of V there correspond values of V' of different sign and *magnitude*—a relatively small one for the direction of the electron stream from the metal into the semi-conductor, and a relatively large one for the opposite direction. Thus the practical rectification coefficient must be always much smaller than the theoretical one given above.

34. Electrical and Thermal Conductivity of Metals (Corpuscular Theory)

The conduction of electricity within a metal (or a semi-conductor) is due to a drift of electrons in the direction of the applied electrical force.

This process can be visualized as similar to some extent to that which we have just been considering, namely, to the passage of electrons through the gap forming the contact between two bodies, these bodies being replaced by the separate atoms, and the width of the gaps between them being reduced to the order of magnitude of 10^{-8} cm.†

† It is interesting to note in this connexion that Faraday was led by his researches on the properties of dielectrics and their relationship with conductors (metals) to an emphatic repudiation of the atomic structure of matter. He argued that since a vacuum acted like a dielectric, there would be no conduction of electricity between the atoms of a material body so long as they are separated from each other by gaps of empty space. In spite of the fact that his conclusion was erroneous and hampered for a long time the progress of our understanding of the nature of the electric current, Faraday

The analogy is, however, incomplete for the following reason: In dealing with the transition of an electron from one metallic body to another (or to a semi-conductor) we assumed that the normal distribution of velocities between the electrons in each body (corresponding to Fermi's or Maxwell's law) remained unperturbed, i.e. the same as in the equilibrium state. This implies that the additional kinetic energy V' acquired by each electron effecting the transition in the positive direction was immediately got rid of, i.e. transferred to the atoms or to the body as a whole in the form of *heat motion*. The finite value of the conductivity of a contact or of its resistance (V'/I) is intimately connected with this transformation of 'electrical energy' into heat.

The conditions which secure such a transformation are obviously not present in the case of a 'contact' between the separate *atoms*. Instead of losing the energy V' acquired in a transition from one of them to the next, the electron can in this case preserve it, increasing it by the same amount in the following transition, and in this way accumulate energy indefinitely. This means that the electron can move in a metal without any resistance, or, in other words, that the electrical conductivity of a metal (or indeed of any other material body) can be infinite.

For this condition to be realized it is necessary that the atoms of the body should not be able to increase the energy of their heat motion at the cost of the electrons. It would be realized, for instance, if they were forced to remain in definite fixed positions, which need not necessarily form a regular crystal lattice.

We have discussed already in § 31 the character of the motion of the electrons in an ideal lattice (in the absence of external electric force). If the atoms are arranged in a fixed but not quite regular way, the 'electron wheels' will be occasionally deviated from their straight course, without, however, any change in the energy of the motion, i.e. the magnitude of the velocity. If, finally, the atoms are not kept fixed, but are allowed to oscillate about certain equilibrium positions—as is actually the case in a real material body—there will be an exchange of energy between the electrons and the atoms or ions, which will serve to maintain the equilibrium distribution for both of them so long as the body has a constant temperature and is not acted upon by external electric forces. Now in the presence of such forces (acting in a closed

should be credited with the true appreciation of the fundamental difficulty connected with the problem of the conduction of electricity in a material body from the point of view of classical mechanics.

circuit) this equilibrium distribution will be destroyed in view of the fact that the electrons will acquire an additional (drift) velocity in the direction of the applied force. Their mutual action with the atoms will tend, however, to restore the distribution by a steady transfer of the larger part of the kinetic energy accumulated by the electrons through the action of the field to the atoms, resulting in a gradual increase of the body's temperature.

In considering the motion of the electrons in a conducting body under the influence of an external electric force, we must thus take into account not only the changes of their momentum, due to a lack of regularity in the crystal lattice of the body, but also the associated changes of the energy, due to the mobility of the atoms. These changes are usually referred to as *collisions* between the electrons and the atoms, and visualized as similar to the collisions between the particles of a gas under the tacit assumption that the irregular changes in the direction of the velocity (or momentum) are always accompanied by an adequate change of the energy just compensating the influence of the external field from one collision to the next one. This assumption is equivalent to the identification of the mean time between two successive collisions with the time of 'relaxation', which is necessary for a practical re-establishment of a slightly perturbed equilibrium distribution.

The mutual action of the electrons in the presence of an external force driving them in a definite direction can be described in terms of a self-consistent field just as well as when there is no such force. We thus avoid the difficulties which confront us if we wish to make a clear picture of the way in which an electron can pass from one atom or ion to the next one, without explicit attention to the fact that the latter must be vacant in order to accommodate it.

The conception that the 'free' electrons in a metallic body can be dealt with as the particles of an ideal gas suffering collisions with the atoms of the body was introduced by Drude. This conception was originally very vague and naive, inasmuch as the question about the number of free electrons and the length of their mean free path was left open, while their motion was identified with that of the particles of an ordinary gas subjected to no forces other than those responsible for the collisions. It was assumed further that the velocity distribution among the electrons is determined by Maxwell's law, their average kinetic energy being thus equal to $\frac{3}{2}kT$.

As has been shown above, this classical conception of the electron gas must be substantially modified in order to be taken over into the

quantum theory. In the first place the number of free electrons in unit volume is here fixed, and of the same order of magnitude as that of the atoms, while their motion can be represented on the corpuscular picture not as a force-free translation but as a rolling of a wheel along an atomic chain. In the second place the Maxwell distribution law is replaced by that of Pauli and Fermi. With these amendments the electron gas theory can be successfully applied, as has been shown by Sommerfeld, not only to such phenomena as do not involve the effects of collisions but also to such phenomena as depend upon collisions, such as, for instance, the conduction of electricity and of heat in a metallic body.

It is necessary, however, in this case, to introduce an empirical parameter characterizing the *mean free path* of an electron, i.e. the average distance l (reckoned from any point on the electron's path) which it will travel without suffering a collision. It must be further assumed (as was done by Drude) that a collision results in a completely irregular change of the direction of velocity and in the restoration of its average magnitude if the latter was altered owing to the action of external forces. This means that the surplus kinetic energy communicated to the electron by these forces must on the average be wholly transferred to the atoms involved in the collision. This surplus energy is extremely small compared with the average kinetic energy of an electron, so that the time τ during which it is accumulated can be identified with the ratio l/v, where v is the 'unperturbed' velocity of the electron.

Under these conditions the electrical conductivity can be very easily calculated—most simply in the following way pointed out by Drude. Under the influence of an external (homogeneous) electric field E an electron moving normally (in the absence of this field) with the velocity v will acquire between two successive collisions an additional velocity in the direction of E equal to $\epsilon E\tau/m = \epsilon El/(mv)$. The mean value of this additional velocity for a given value of the normal one v is thus equal to $\epsilon El/(2mv)$. Averaging it for all values of v and multiplying by the number n of electrons in unit volume and the charge ϵ of an electron, we obtain the quantity of electricity transported by the electrons per unit time and area in the direction of the applied field, i.e. the current density

$$j = \sigma E,$$

where σ is the specific conductivity. We thus find for the latter the expression

$$\sigma = \frac{\epsilon^2 nl}{2m}\overline{v^{-1}}. \tag{175}$$

Since in a metallic body the velocity distribution of the electrons is practically independent of the temperature, we can calculate the average value of $1/v$ or $1/(mv) = 1/g$ for $T = 0$ from the general formula

$$\overline{F(g)} = \int_0^g F(g) g^2 \, dv \Big/ \int_0^{g_0} g^2 \, dg,$$

where g_0 is the maximum value of g at $T = 0$. We thus get

$$\overline{\left(\frac{1}{g}\right)} = \frac{3}{2}\frac{1}{g_0},$$

and consequently

$$\sigma = \frac{3}{4}\frac{n\epsilon^2 l}{g_0} = \frac{3}{4}\left(\frac{8\pi}{3}\right)^{\frac{1}{3}}\frac{n^{\frac{2}{3}}\epsilon^2}{h^2} l. \tag{175a}$$

With $n \cong 10^{22}$ cm.$^{-3}$ satisfactory agreement with experimental data is obtained if for ordinary temperatures l is equal to about 10^{-6} cm.

As is well known, metallic bodies are distinguished not only by their electrical conductivity, but also by their exceptionally high thermal conductivity. In fact, the two conductivities are found to be proportional to each other. This result, which forms the essence of the Wiedemann-Franz law, was interpreted by Drude on the basis of his theory, which can be readily applied to the calculation of the thermal conductivity, if one assumes that in a metallic body the heat energy as well as the electric charge is transported by the free electrons. Following the scheme of the kinetic theory of gases we can define the heat conductivity by the formula

$$\kappa = Dnc_e = \tfrac{1}{3}l\bar{v}nc_e, \tag{176}$$

where c_e is the specific heat per electron and $D = \frac{1}{3}l\bar{v}$ the diffusion coefficient. It should be noticed that, according to the classical statistics which forms the basis of Drude's theory, c_e is equal to $\frac{3}{2}k$. Taking the ratio κ/σ and identifying the ratio $\bar{v}/\overline{v^{-1}}$ with $\overline{v^2}$ (which of course is not correct) we get on this theory, since $m\overline{v^2} = 3kT$,

$$\frac{\kappa}{\sigma} = 3\left(\frac{k}{\epsilon}\right)^2 T. \tag{176a}$$

This formula was derived by Drude and is in excellent agreement with the experimental facts. If in calculating $\bar{v}$ and $\overline{v^{-1}}$ we use Maxwell's distribution law, we get, putting $m/(2kT) = \alpha$,

$$\bar{v} = \int_0^\infty e^{-\alpha v^2} v^3 \, dv \Big/ \int_0^\infty e^{-\alpha v^2} v^2 \, dv = \frac{1}{\sqrt{\pi\alpha}}$$

and $$\overline{v^{-1}} = \int_0^\infty e^{-\alpha v^2} v\, dv \Big/ \int_0^\infty e^{-\alpha v^2} v^2\, dv = \sqrt{\frac{\pi}{\alpha}},$$

whence $$\frac{\bar{v}}{\overline{v^{-1}}} = \frac{1}{\alpha} = \frac{2kT}{m},$$

and consequently $$\frac{\kappa}{\sigma} = 2\left(\frac{k}{\epsilon}\right)^2 T. \qquad (176\,b)$$

This formula was obtained by Lorentz by a more complicated method; so far as the numerical cofficient is concerned it is in poorer agreement with the experimental data than Drude's formula (176 a).

It is remarkable that the replacement of classical statistics by that of Pauli and Fermi does not substantially alter Drude's result (which is certainly quite incorrect from the point of view of its theoretical foundations). We have in fact in this case, according to (143 a),

$$c_e = \frac{\pi^2 k^2 T}{2W_0} = \frac{\pi^2 k^2 T}{m v_0^2},$$

and further $$\overline{v}/\overline{v^{-1}} = \int_0^{v_0} v^3\, dv \Big/ \int_0^{v_0} v\, dv = \tfrac{1}{2} v_0^2,$$

whence it follows that $$\frac{\kappa}{\sigma} = \frac{\pi^2}{3}\left(\frac{k}{\epsilon}\right)^2 T. \qquad (176\,c)$$

This formula is due to Sommerfeld (1927). It is interesting to note that the temperature T enters the result in the quantum statistics not through the kinetic energy of the electrons (as in the Drude-Lorentz theory) but through the heat capacity c_e.

It should be mentioned that the derivation of Wiedemann-Franz law in the classical electron theory of Drude and Lorentz involved a contradiction with the experimental facts about the heat capacity of the metals. If in order to account for the abnormally high heat conductivity of metals we assume that it is due to free electrons, then the number of the latter must be of the same order of magnitude as the number of atoms. But in this case on the classical theory they should contribute markedly to the specific heat of the metal, increasing the latter by the amount corresponding to a classical ideal gas, i.e. $\frac{3}{2}k$ per electron. If the number of electrons was just equal to the number of atoms, we would thus obtain 9 calories per gramme-atom, instead of the 6 calories which correspond to the law of Dulong and Petit. This contradiction between the role of the free electrons in the specific heat of the metal and in the conduction of heat is very simply removed in the quantum theory of Sommerfeld.

35. Wave-mechanical Theory of Collisions in an Electron Gas

Sommerfeld's theory of the electrical and thermal conductivity of a metal is incomplete, inasmuch as it does not attempt to determine the mean free path of the electrons l. Now so long as this quantity remains indeterminate we can say nothing about the absolute magnitude of σ (or κ) and its dependence on the temperature (all the other quantities entering in the expression (175) for σ being practically, if not wholly, independent of T).

A rational determination of the mean free path l can be carried out in an elementary way with the help of wave mechanics on the basis of the following argument which can be considered as a new application of the idea of the wave-corpuscular parallelism.†

Let us imagine a stream of electrons moving in the direction of the x-axis through some material (metallic) body. The mean free path of an electron can be defined as the denominator of the expression dx/l which gives the probability that it will be deflected from the original direction of motion on a distance dx along its path. Thus if I electrons are passing in the x-direction through unit area at some point x, the number of electrons that will suffer a deflexion between x and $x+dx$ will be equal to $I\,dx/l$, and this can be equated to the decrease in the number of electrons passing without deflexion the distance dx. We thus get $-dI = I\,dx/l$, or

$$I = I_0\, e^{-x/l}. \tag{177}$$

This is the well-known formula of Clausius (originally derived for a beam of molecules moving through a gas). The factor $e^{-x/l}$ can be interpreted as the probability that an electron will pass without deflexion a distance equal to or larger than x (the average value of x being equal to l).

Now a perfectly similar formula describes the decrease of the intensity of a beam of light rays travelling through a body which absorbs or scatters them. It can be applied just as well to the propagation of cathode rays considered from the point of view of the wave theory, the intensity of the waves, measured by the product of the quantity $\psi\psi^*$ by the velocity, being the wave-mechanical equivalent of the probable number of electrons I crossing unit area in unit time. As has been pointed out above, a deflexion of an electron (i.e. a change of its momentum) in order to be effective in the creation of electrical resistance must be combined with a change of kinetic energy. The wave-

† Introduced by the author in 1927 (and independently by Sommerfeld without, however, any quantitative development).

mechanical equivalent of this combined process (which can take place just as well in the case of light) is called a 'combined scattering' of the waves, i.e. an irregular change of their wave vector **k**, accompanied by a change (on the average a decrease) of the frequency ν.† Leaving this change of frequency aside for a while we can measure the amount of scattering by a certain extinction or scattering coefficient μ, which determines the decrease of intensity I as a function of the distance x travelled, by the equation

$$I = I_0 e^{-\mu x}. \tag{177 a}$$

Comparing this equation with (177) we get

$$l = \mu^{-1}, \tag{177 b}$$

which means that the mean free path of the electrons can be defined in terms of the wave theory as the reciprocal of the extinction (scattering) coefficient of the corresponding waves.

For cathode waves with a large wave-length, i.e. large compared with the distance between the neighbouring atoms of the body under consideration, this scattering must be quite similar to the Tyndall effect for ordinary light waves in transparent bodies. As a consequence of the irregular variations or fluctuations of density due to the heat motion, these bodies acquire a virtual inhomogeneity, which is revealed by the appearance of a diffusely scattered light on the one hand and by the extinction of the primary light rays according to the equation (177 a) on the other.

This equation can also be applied to the scattering of X-rays whose wave-length is of the same order of magnitude as, or still smaller than, interatomic distances. In this case, however, the scattering can be ascribed directly to the discontinuous (i.e. atomic) structure of the body, which itself will be revealed as an inhomogeneity so long at least as the atoms are not arranged with absolute regularity.

As we have seen above, the fastest electrons in a metallic body (at the temperature of absolute zero) correspond to wave-lengths of the same order of magnitude as the interatomic distances. We have, for instance, in the case of sodium, $\lambda_{\min} = 5{\cdot}4 \times 10^{-8}$, which is about twice the nearest distance between the atoms. Such electron waves should accordingly be compared in respect to their scattering with soft X-rays rather than with ordinary light. For the sake of simplicity we shall,

† This 'combined' scattering in the case of light is usually called the Raman effect, whereas the scattering unaffected or practically unaffected by a change of frequency which is the usual type observed with light (or X-rays) is called the Tyndall effect, or more commonly, but incorrectly, Rayleigh scattering.

however, treat the scattering of electron waves in a metal as analogous to the ordinary Tyndall effect, neglecting the somewhat exceptional condition of the shortest waves that come into question.

Let μ_1 denote the relative intensity of light waves or cathode rays scattered in all directions by a single atom under the influence of primary waves of unit intensity. Let, further, N_a denote the number of atoms in a given volume V small compared with the wave-length. So far as under these conditions the secondary waves scattered by all the atoms in V are in phase with each other, their amplitudes will be arithmetically added, the resulting amplitude being proportional to N_a and the total intensity being equal to $\mu_1 N_a^2$. We are concerned here only with that part of this intensity which corresponds to an irregular variation of N_a (in time and in space, i.e. from one volume element V to another) about its average value $\overline{N_a}$, and which will appear in the form of irregularly scattered radiation (the rest combining into regularly transmitted or refracted waves). This part is equal to

$$\mu_1 \overline{(N_a - \overline{N_a})^2} = \mu_1 \overline{(\Delta N_a)^2}.$$

Denoting the concentration of the atoms in the volume V, i.e. the ratio N_a/V, by n_a and its average value by $\overline{n}_a$, we thus get for the total intensity scattered by the volume element V of the body the expression $\mu_1 \overline{(\Delta n_a)^2} V^2$. Multiplying this expression by the number of such elements in unit volume, i.e. by $1/V$, we obtain the scattering coefficient of the equation (177 a):

$$\mu = \mu_1 \overline{(\Delta n_a)^2} V. \tag{178}$$

It should be mentioned that for very short cathode waves or X-rays whose wave-length is of the same order of magnitude or still smaller than the interatomic distances we obtain instead of (178) the simple formula

$$\mu = \mu_1 \overline{n_a}. \tag{178 a}$$

This corresponds to a complete incoherence of the secondary waves scattered by the single atoms, and holds, strictly speaking, only in the case of a lack of absolute regularity in the structure of the atomic lattice (there being for X-rays just as for ordinary light no irregular scattering at the temperature of absolute zero). For ordinary light, or for cathode waves of a relatively large wave-length, the decrease of μ with decrease of the temperature follows immediately from equation (178), since the fluctuations of the concentration Δn_a so long as they are due to heat motion, vanish for very low temperatures.

It is interesting to note that the two equations (178) and (178 a)

would coincide for a *gas*. In fact, since the atoms are here independent of each other (in regard to their motion or position), we can calculate the average value of the quadratic fluctuation $(N-\overline{N})^2$ by means of the formula (137 b), § 28 (although it refers to the phase-space and not to ordinary space). This gives $\overline{(\Delta N_a)^2} = \overline{N_a} = \overline{n_a} V$ or $\overline{(\Delta n_a)^2} = \overline{n_a}/V$, so that equation (178) is reduced to (178 a).

For a pencil of relatively slow cathode rays travelling through a gas, the mean free path can be determined by the well-known formula

$$l = \frac{1}{\pi a^2 \overline{n}_a},$$

where a denotes the ordinary 'radius' of an atom, which can be considered as a constant of the order of magnitude of 10^{-8} cm. Replacing l by $1/\mu = 1/(\mu_1 \overline{n}_a)$ we thus get

$$\mu_1 = \pi a^2.$$

The coefficient μ_1 can be defined accordingly as the effective cross-section of an isolated atom for a collision with a relatively slow electron.

The expression (178) for μ can also be written in the form

$$\mu = \pi a_1^2 \overline{n}_a,$$

corresponding to an effective atomic radius a_1 given by

$$a_1^2 = a^2 \frac{\overline{(\Delta n_a)^2} V}{\overline{n}_a}. \tag{179}$$

We can thus use for a solid or liquid metal the same picture of the collision process as for a metallic vapour if the individual cross-section of an isolated atom is replaced by an 'apparent' cross-section πa_1^2, which is in general much smaller and depends upon the temperature.

Its dependence upon the temperature can be obtained as follows:

Let us consider not a definite volume element of the body but a definite number of atoms N which in the equilibrium state occupy the volume $V = \overline{V}$, so that $N/V = \overline{n}_a$. The density fluctuations will be then replaced by fluctuations of this volume. These fluctuations (ΔV) are accompanied by a change of energy $\Delta U = \frac{1}{2}\kappa(\Delta V)^2/V$, where κ is the modulus of compressibility of the body. The probability that V will be changed spontaneously, i.e. as a consequence of the heat motion, by the amount $\Delta V = \xi$ is proportional to $e^{-\Delta U/kT} = e^{-\alpha\xi^2}$, where $\alpha = \kappa/(2VkT)$. We thus get for the average value of $(\Delta V)^2$ the expression

$$\overline{(\Delta V)^2} = \int_{-\infty}^{+\infty} e^{-\alpha\xi^2} \xi^2 \, d\xi \Big/ \int_{-\infty}^{+\infty} e^{-\alpha\xi^2} \, d\xi = \frac{1}{2\alpha} = \frac{VkT}{\kappa}.$$

Now so long as N is constant we must have

$$\overline{n}_a \Delta V = -\overline{V}\Delta n_a,$$

and consequently

$$\overline{(\Delta V)^2} = \overline{(\Delta n_a)^2}\frac{V^2}{(\overline{n}_a)^2};$$

whence it follows that

$$\frac{\overline{(\Delta n_a)^2}V}{\overline{n}_a} = \frac{\overline{n}_a kT}{\kappa}. \tag{179 a}$$

Comparing this with (179) we get

$$\left(\frac{a_1}{a}\right)^2 = \frac{\overline{n}_a kT}{\kappa}. \tag{179 b}$$

This gives for sodium at room temperatures ($\overline{n}_a \cong 2{\cdot}5.10^{22}, \kappa = 6{\cdot}5.10^{10}$, $T \cong 300$)

$$\left(\frac{a_1}{a}\right)^2 \cong 1{\cdot}5.10^{-2},$$

that is, an apparent radius a_1 which is about ten times smaller than the effective radius of an individual atom.†

It should be mentioned that this apparent radius can be identified with the amplitude of the heat oscillations of the atom about its equilibrium position. In fact, by the definition of the modulus of compressibility κ (see formula (148 a), § 30), the expression $\kappa a_1^2/(n_a a^2)$ represents approximately the potential energy of an atom when it is removed from its equilibrium position at a distance a_1. Equating this energy to kT—which gives the order of magnitude of the average value of the potential energy at the temperature T—we obtain the relation (179 b).

Substituting (179 a) in (178), and replacing μ by its reciprocal, we obtain the following expression for the mean free path of the conduction electrons in a metallic body:

$$l = \frac{\kappa}{\pi a^2 n_a^2 kT}. \tag{180}$$

With this expression for l, the electric conductivity as determined by equation (175 a) turns out to be *inversely proportional to the absolute temperature*—a result which is in agreement with the observed facts for ordinary and higher temperatures (below the fusing-point). For $T = 300$, l is numerically equal to 10^{-6} cm.; as pointed out before, this leads to values of the conductivity which are of the correct order of magnitude.

† The mean free path l must accordingly be about 100 times larger than the interatomic distances—in agreement with the conclusion derived from the value of the conductivity.

The above theory is further supported by two well-known experimental facts, namely, the abnormally high electrical resistance of metals containing impurities of any kind, and the increase of resistance in the transition from the solid (crystalline) to the liquid state. According to the wave-mechanical conception the impurities must communicate to the metal the properties of an unhomogeneous medium in regard to the propagation of electron waves even at the temperature of absolute zero, and thus increase the scattering coefficient of these waves μ by an amount $\Delta\mu$ which is proportional to the concentration of the impurities and which is practically independent both of their chemical nature (they may be in particular atoms of a still better conducting metal) and of the temperature. This is exactly what is observed experimentally with regard to the specific resistance of metals containing impurities (Mathiessen's rule).

It should be mentioned that the *relative* importance of the impurities increases as the temperature decreases, and the spontaneous inhomogeneity due to the heat motion of the metal dies out.

As to the increase of electrical resistance at the fusion-point (about twofold), it seems a quite obvious consequence of the disappearance of that regularity which is characteristic of a crystalline substance. It must be kept in mind, however, that this lack of regularity would not affect the scattering of the electron waves if they were long enough compared with the interatomic distances (as is the case with ordinary light). The increase of the scattering coefficient should then be wholly attributed to a decrease of the compressibility modulus κ. Such a conclusion does not seem, however, to check quantitatively with the experimental data. The increase of resistance at the fusion-point thus serves to show that in reality we have to deal with electron waves whose length approaches in order of magnitude the interatomic distances.

The scattering of such short electron waves should take place in a way similar to the scattering of soft X-rays rather than of ordinary light, the corresponding part of the scattering coefficient μ being dependent upon the temperature to a much smaller extent than the part corresponding to electron waves of greater length. Applying Debye's theory of the influence of heat motion on the scattering of X-rays in crystals to short electron waves, one gets for μ an expression of the form

$$\mu \sim \left(1 - \frac{1 - e^{-\gamma T}}{\gamma T}\right),$$

which reduces to $\mu \sim T$ if $\gamma T \ll 1$. It should be mentioned, however,

that it is valid for relatively high temperatures only, for which the energy of the heat oscillations is proportional to T. At lower temperatures their energy decreases much more rapidly with decrease of the temperature. One gets accordingly a much more rapid decrease of electrical resistance, especially for long electron waves, than that which follows from the expression (180) for the mean free path (see next section).

36. 'Collectivistic' Treatment of the Motion of Electrons in a Metal

In the preceding considerations we have left out of account the fact (pointed out by L. Brillouin) that the scattering of the electron waves—whether long or short—must conform to Pauli's exclusion principle. Any scattering processes that would lead to a violation of this principle must accordingly be barred. It may seem at first sight that this restriction must greatly reduce the value of the scattering coefficient obtained above, even for ordinary temperatures, and thus completely destroy the quantitative agreement between our theory and experiment. As has been pointed out, however, in § 20, the maintenance of Pauli's principle for a system of electrons subject to external forces only (their interaction being replaced by a self-consistent electric field) is secured automatically even with classical mechanics, in virtue of Liouville's theorem. This circumstance can be more clearly understood if, instead of following the career of separate electrons, we consider the resulting change in the distribution of all the electrons among the different states of motion in the crystal lattice of the metal. A transition of one of the electrons, as a consequence of a collision process, into a state which was previously occupied by another electron must take place simultaneously with a transition of this second electron to some new state. We meet here the same difficulty as that which is met with when instead of the 'lattice states' we consider 'atomic states'—an electron being then able to pass from one atom to another only if the latter is simultaneously left vacant to accommodate it. This difficulty, however, is only apparent, since all the electrons are moving simultaneously, their mutual adjustment being a simple corollary of the laws of motion.

This mutual adjustment will effect a partial mutual compensation of the individual transitions, so far as the distribution of the electrons between different individual states is concerned. If the free electrons in a metal are dealt with 'collectively' and not 'individually', all those collision processes which are mutually compensated can be left out of

account, the net result being determined by the non-compensated unidirectional processes only. Processes of this kind must take place only so far as the stationary (equilibrium) distribution of the electrons among the individual states of motion in the crystal lattice of the metal at $T = 0$ is disturbed by the action of the external electric field and by the heat motion of the atoms. It is further clear that the change of statistical distribution will practically be limited to the boundary region of the normal Fermi distribution, i.e. to the top of the normally occupied lower portion of the phase-space skyscraper, where the electrons can move individually, without any compensating processes, to vacant cells lying higher up.

From this 'collectivistic' point of view the action of an electric field on the free electrons in a metal can be considered to consist not in a change of the motion of the separate electrons compared with their motion in the absence of the field, but in a change of their *distribution* between the *unperturbed* states of motion, that is, in a change of the number of electrons associated with these states. We obtain in this way a perturbed statistical distribution, which can be summarized as follows: In the absence of the field the electrons will be distributed in pairs moving freely with the same energy *in opposite* directions and thus giving no resulting current. (This 'doubling' is of course quite independent of the spin doubling.) The electric field will tend to increase the number of electrons moving in the positive direction at the cost of those which move in the negative direction, this redistribution being quite similar to that which is produced by a magnetic field on the electron spins (§ 30). Like the magnetic field, the external electric field will only affect the electrons in the highest (occupied) states, causing the pairs of electrons moving in opposite directions with the same energy to part, and inducing that partner which moved originally in the negative direction to pass to a formerly vacant cell with *higher* energy, corresponding to a motion in the positive direction. The only difference between the action of the magnetic field on the orientations of the spins and the action of the electric field on the directions of motion of the electrons is that, so long as the electric field is acting in a closed circuit (which means a non-conservative motion), its effect will steadily increase with the time, unless balanced by the effect of the heat motion which will tend to restore the original equilibrium distribution. As a result a current will be obtained which will increase with the applied field and decrease with increase of the temperature.

If the electrons are strongly bound to the individual atoms their 'lattice states' of motion will form, as we know, narrow bands separated by wide gaps, which correspond to selective reflection. Under these conditions the conductivity will be very small whatever the number of the corresponding electrons, because of the smallness of the transition probability from one atom to the next one. It is interesting to note that it must reach a maximum value when the number of electrons corresponding to the whole band of lattice states is equal to just one-half of the number of these states. If the whole band is filled the electrons will be unable to pass to states lying higher up in the phase-space skyscraper—at least under the influence of moderate electric fields—for these states are separated from the original ones by a long range of floors corresponding to forbidden motion (selective reflection).

This explains why the inner electrons, though able in principle just as the outer ones to pass from one atom to the next (although with an extremely small probability) do not play any role whatever in the conduction of electricity in a metallic body. It serves also to explain the power of many non-metallic bodies to become conductive under the influence of light (inner photo-electric effect) or heat (semi-conductors). In both cases some of the electrons are thrown from a state in a narrow band—filled or unfilled—into a state lying much higher up and corresponding to a relatively rapid motion through the lattice, i.e. described by a wave function $\psi = C(\mathbf{r})e^{i\gamma\cdot\mathbf{r}}$ with but a slightly variable periodicity factor $C(\mathbf{r})$.

We shall not pause here to undertake a quantitative development of this 'collectivistic' theory of the motion of the electrons in a solid body. This rather complicated theory which has been worked out by F. Bloch and by R. Peierls will be examined later on. An important part of this theory is concerned with giving an adequate description of the heat motion of the atoms in a solid body and of the interaction of this motion and that of the electrons, in terms of the exchange of momentum and energy between the electrons and the elastic or sound waves representing the heat motion in a solid according to Debye's theory. It is possible to give a very simple and vivid interpretation of the corresponding results, using the 'individualistic' description of the motion of the electrons, as we have done before, and introducing the fictitious but sometimes very convenient conception of 'phonons', i.e. sound or heat quanta associated with the elastic (or acoustical) waves which serve for the description of heat motion in a solid body.

37. The Theory of a 'Phonon Gas' (Mechanical Heat in a Solid) and of its Interaction with the Electron Gas

The acoustical waves representing the heat motion in a solid body, as defined by their wave-number vectors $\mathbf{k}$ (which determine the wavelength $\lambda = 1/k$ and the direction of propagation), are exactly similar to the electron waves which describe the motion of the free electrons in the same body at the temperature of absolute zero, if the number of these electrons is assumed to be twice as large as the number of atoms. To each of these electron waves there correspond three acoustical waves with exactly the same vector $\mathbf{k}$, one of them being a longitudinal wave, and the two others transversal ones with mutually perpendicular directions of vibration.

This coincidence follows at once from the fact that the two types of waves—electron and sound—considered as standing waves, are specified by practically the same boundary conditions (the surface of the body being a nodal surface), while in both cases the number of different waves of the same sort (or 'normal vibrations') is equal to the number of atoms N. This is a consequence of Pauli's principle for the electron waves, and of the fact that the number of normal modes of vibration must be equal to the number of degrees of freedom of all the N atoms constituting the body for the sound waves, while finally in both cases the waves are limited to the long-wave portion of the spectrum.

This spectrum must therefore extend from a practically vanishing value of the wave-number k up to the same maximum value k_0 which is determined by the formula†

$$\frac{4\pi}{3}k_0^3 = \frac{N}{V} = n_a. \qquad (181)$$

If we replace k_0 by g_0/h and n_a by $\frac{1}{2}n$, we obtain the relation (142 a) of § 30. In a similar way the number dN of waves (normal vibrations) for which the magnitude of the vector $\mathbf{k}$ (irrespective of its direction) lies within assigned limits is given by the equation

$$dN = 4\pi V k^2\, dk. \qquad (181\,a)$$

In Debye's theory each standing acoustical wave is dealt with as a harmonic oscillator with a definite frequency $\nu_l = c_l k$ for the longitudinal wave and $\nu_t = c_t k$ for the two transverse ones, c_l and c_t being the corresponding velocities of propagation. The motion of each oscillator is further 'quantized' in the sense that its energy is assumed to

† Cf. § 11, Ch. III, where the number of normal modes of vibration was determined directly for the special case of a rectangular box.

be equal to an integral multiple of the corresponding quantum $h\nu$, i.e. $h\nu_l$ or $h\nu_t$. According to quantum mechanics the integral numbers must be replaced by half-integral ones. The number of energy quanta associated with each wave (mode of vibration) remains arbitrary, while the probability that at a given temperature T the harmonic oscillator with a frequency ν shall have an energy $W_n^{(\nu)} = nh\nu$ is determined according to the classical statistics by Boltzmann's formula

$$P_n = \text{const. } e^{-nh\nu/kT}. \tag{182}$$

The average value of this energy is thus equal to

$$\overline{W^{(\nu)}} = \frac{\sum_{n=0}^{\infty} nh\nu e^{-nh\nu/kT}}{\sum_{n=0}^{\infty} e^{-nh\nu/kT}} = -h\nu \frac{\partial}{\partial \alpha} \log \sum_{n=0}^{\infty} e^{-n\alpha} = +h\nu \frac{\partial}{\partial \alpha} \log(1-e^{-\alpha}),$$

where $\alpha = h\nu/kT$, that is, to

$$\overline{W^{(\nu)}} = \frac{h\nu}{e^{h\nu/kT}-1}. \tag{182 a}$$

Replacing n by $n+\frac{1}{2}$ in agreement with the wave-mechanical theory of a harmonic oscillator, we easily get instead of (182 a)

$$\overline{W^{(\nu)}} = \frac{h\nu}{e^{h\nu/kT}-1} + \tfrac{1}{2}h\nu. \tag{182 b}$$

The second term represents the 'zero-point energy' of the oscillator, i.e. its energy at the temperature of absolute zero, for which the first term vanishes. For sufficiently high temperatures $\overline{W^{(\nu)}}$ tends to the classical expression kT.

The formulae (182 a) and (182 b) are actually due not to Debye but to Planck, who derived them, in his well-known theory of the thermal radiation, for a material harmonic oscillator bathed in this radiation. Debye applied them subsequently to the oscillators describing the radiation itself and in this way arrived at a complete analogy between the theory of 'radiant heat' and that of a 'mechanical heat', constituted by the heat motion of the atoms in a solid body. This analogy is restricted only by the fact that in a solid body the spectrum of the waves does not extend to infinity (in the direction of increasing wave numbers) but stops at a fixed limit given by equation (181).

Now just as for light and electrons it is possible to associate the acoustical waves with certain particles which we shall call 'phonons', and to replace the study of the heat oscillations forming these waves by the study of the motion of the corresponding 'phonons'.†

† It is not in the least intended to convey hereby the impression that such phonons

Since the motion of a phonon with a given momentum $\mathbf{g} = h\mathbf{k}$ corresponds to a plane harmonic wave travelling in a given direction, we must replace standing sound waves by travelling ones, by whose superposition standing waves can be formed (cf. § 12).

A representation of the momentum of a phonon as the product of its 'mass' and 'velocity' has no physical meaning, one of the two factors being in fact quite arbitrary. The velocity of a phonon might in particular be identified with the velocity of the corresponding waves, in which case its mass would be proportional to the wave number or to the frequency just as in the case of photons.

Another point of similarity between the photons and the phonons is that *for neither is the number fixed*—in contradistinction to electrons—so that phonons can be created and annihilated in various processes analogous to the emission and absorption of light.

Inasmuch as the heat motion of the atoms, represented by the superposition of Debye's sound waves, is replaced by a sort of ideal gas formed by the phonons, the scattering of the electrons (which form another sort of ideal gas) must be ascribed to their interaction with the phonons. This interaction can be visualized either as a process of *collision*, involving a simultaneous change of momentum and energy of the two particles concerned (electron and phonon), or as a process in which a phonon is emitted or absorbed, as it were, by the electron. In both cases the total momentum and total energy must remain the same, i.e. we must have the equations

$$\mathbf{g}''-\mathbf{g}' = h\mathbf{k}'-h\mathbf{k}'', \qquad W''-W' = h\nu'-h\nu'' \tag{183}$$

in the case of an ordinary collision, and

$$\mathbf{g}''-\mathbf{g}' = \mp h\mathbf{k}, \qquad W''-W' = \mp h\nu \tag{183 a}$$

in the case of emission (upper sign) or absorption (lower sign). Here $\mathbf{g}$ and W denote the momentum and the energy of an electron, while $h\mathbf{k}$ and $h\nu$ have the same meaning for the phonons; the primed letters refer to the initial and the double primed to the final state.

It should be mentioned that the collision processes described by (183) are the exact analogy of the collisions between electrons and photons constituting, from the corpuscular point of view, the Compton effect.

Emission and absorption processes of the type (183 a) are, however, impossible with photons, so long as we are dealing with free electrons (the condition for their possibility being that the velocity of the electron

have a real existence. On the contrary, the possibility of their introduction rather serves to discredit the belief in the real existence of photons.

g/m should be larger than the velocity of the sound or light waves ν/k.) The wave-mechanical analysis of the scattering of the electron waves by sound waves shows, however, that this scattering is mainly due to the emission and absorption processes of the type (183) and not to collision processes of the type (183 a). The same result is found in the case of the scattering of *light* in solids, the photons playing here the role of the electrons; the equations (183 a) represent in this case the 'combined scattering' of light or the 'Raman effect' due to the heat vibrations of the atoms.

Let us consider first the collision processes. They can be specified by defining a certain effective radius a or cross-section $\mu_1 = \pi a^2$ representing the target area for a collision between an electron and a phonon. Assuming this quantity to be independent of both W and ν (which need not be the case) we can determine that part of the scattering coefficient of the electrons, which is due to these collisions as a function of the number of phonons in unit volume n_p by the usual formula

$$\mu = \pi a^2 n_p. \tag{184}$$

If emission and absorption processes could be disregarded, the reciprocal of μ would give the mean free path of the electrons l.

The number of phonons corresponding to sound waves of a definite type and lying within a certain range dk of the wave-number spectrum is equal to the product of (181 a) by (182 a) divided by $h\nu$. Neglecting, for the sake of simplicity, transverse waves (which are probably much less effective than the longitudinal ones with regard to the scattering process), we thus get†

$$n_p = 4\pi \int_0^{x_0} \frac{k^2\, dk}{e^{h\nu/kT} - 1},$$

or, putting $k = \nu/c_l$,

$$n_p = \frac{4\pi}{c_l^3} \int_0^{\nu_0} \frac{\nu^2\, d\nu}{e^{h\nu/kT} - 1} = \frac{4\pi}{c_l^3} \left(\frac{kT}{h}\right)^3 \int_0^{T_0/T} \frac{\xi^2\, d\xi}{e^{\xi} - 1}. \tag{184 a}$$

For temperatures which are high enough compared with Debye's 'characteristic temperature' $T_0 = h\nu_0/k$ we can put with sufficient approximation $e^{h\nu/kT} - 1 \cong h\nu/kT$, which gives

$$n_p = \frac{kT}{h} \frac{4\pi}{2c_l^3} \nu_0^2 = \tfrac{3}{2} kT \frac{4\pi}{3} k_0^3,$$

† The factor k in the product kT (Boltzmann's constant) should not be confused with the wave number denoted by the same letter in $k^2\, dk$.

or, according to (181),

$$n_p = \frac{3}{2}\frac{kT}{h\nu} n_a = \frac{3}{2}\frac{T}{T_0} n_a. \tag{184b}$$

Substituting this expression in (184) we obtain the same dependence of the scattering coefficient on the temperature as that which was derived in the preceding section for the case of high temperatures. For the opposite extreme of very low temperatures we have approximately

$$n_p = \frac{4\pi}{c_l^3}\left(\frac{kT}{h}\right)^3 \int_0^\infty \frac{\xi^2\, d\xi}{e^\xi - 1} = \text{const.}\, T^3. \tag{184c}$$

Similar results are obtained on the basis of the emission and absorption processes. To see this we must take into account the fact that in the equilibrium state at a given temperature T the number of absorptions must be equal to the number of emissions (there being a detailed balance between processes of opposite kinds). Now, since the probability that an electron will absorb a phonon is obviously proportional to the concentration of the latter, the same must be true about the probability of emission, so that the probability of the two processes referred to a unit path of the electron must be proportional to the concentration of the phonons—just as for ordinary collisions.

If instead of following the career of individual electrons we consider their distribution in the velocity or momentum space, we can trace the effect of the accelerating field in a steady increase of the number of electrons moving in the direction of the applied force (with a kinetic energy somewhat larger than W_0) at the cost of electrons which were initially moving in the opposite direction. The scattering (or absorption-emission) processes will oppose this shift, tending to restore the equilibrium distribution. This struggle between the two tendencies will result in an electric current proportional to the electric field and inversely proportional to the concentration of the phonons. Besides, however, this concentration, we must take into account the *efficiency* of the scattering processes in restoring the equilibrium distribution—a circumstance that has been left aside hitherto. We thus see that the scattering coefficient μ or the mean free path $l = \mu^{-1}$ does not fully determine the electrical resistance, but that for a correct determination of the latter it is necessary to multiply μ by a certain coefficient, which gives the average efficiency of the scattering processes in discharging from the electron the additional momentum and energy which it is accumulating through the action of the field. Now according to the

equations (183) and (183a) the momentum transferred by an electron to a phonon is of the order of magnitude of $\pm\hbar k$. In averaging over all directions of $\mathbf{k}$ we obtain for the mean value of the momentum transferred in the direction of the resulting current a quantity proportional to $\overline{(\hbar k)^2}$ or $\overline{\nu^2}$. For sufficiently high temperatures ($T \gg T_0$) this quantity (which is the measure of the effectiveness of the scattering processes) is independent of T and practically coincides with ν_0. In the opposite case ($T \ll T_0$) it is approximately equal to the square of that frequency ν_T for which the function $\frac{\nu^2}{e^{\hbar\nu/kT}-1} \cong \nu^2 e^{-\hbar\nu/kT}$ (number of phonons per unit frequency interval) reaches its maximum value. This frequency is easily seen to be proportional to T (just as for light). We thus get for low temperatures an efficiency coefficient which is proportional to T^2. Since the scattering coefficient in this case is itself proportional to T^3, we see that at very low temperatures the specific electrical resistance ρ of a metal must be proportional to T^5. This result has been obtained by Bloch and by Peierls. It seems to be supported by the experimental facts. For intermediate temperatures the electrical resistance of metals can be expressed by an empirical formula discovered by Gruneisen, $\rho = \text{const.}\, cT$, where c is the specific heat at the temperature T. According to the preceding theory we should expect a somewhat different dependence, though of the same general character. In fact the scattering coefficient μ, and consequently the specific resistance, must be proportional, according to this theory, to the expression (184a) multiplied by an efficiency factor changing from T^2 (for $T \ll T_0$) to a constant (for $T > T_0$), whereas the heat capacity c is equal to the differential coefficient with respect to T of the energy of all the phonons, that is, to

$$W_l = \frac{4\pi}{c_l^3} \int_0^{\nu_{ol}} \frac{h\nu^3\, d\nu}{e^{\hbar\nu/kT}-1} + \frac{8\pi}{c_t^3} \int_0^{\nu_{ot}} \frac{h\nu^3\, d\nu}{e^{\hbar\nu/kT}-1}, \tag{185}$$

if phonons corresponding to transversal sound waves are also taken into account.

In view of the rather crude character of the whole theory we should hardly expect, however, to find an exact quantitative agreement of its results with the experimental data.

It is interesting to note that the heat conductivity of non-metallic bodies can be determined on the basis of the phonon theory by means of the formula (176) if the velocity v is identified with the velocity of

sound, and the product nc_e with the total heat capacity. We thus get for ordinary temperatures a heat conductivity which is inversely proportional to the absolute temperature—in agreement with the experimental data on crystals of dielectrics.

38. The Theory of a Photon Gas (Thermal Radiation)

We shall now apply the general theory to another special case of particular importance, namely, to a 'photon gas', or, in other words, to *radiation* enclosed in a box with perfectly reflecting walls and supposed to be in statistical equilibrium. Such radiation is called thermal, or 'black body' radiation.

According to (139) and (139 a) the number dN of photons of both types of polarization, whose momenta g lie between g and $g+dg$, is given by the formula

$$dN = \frac{8\pi V}{h^3} \frac{g^2\, dg}{e^{\mu(W-W_0)}-1}, \tag{186}$$

where μW_0 stands for $-\lambda$. It has been shown above that for sufficiently low temperatures the quantity W_0 can be identified with the smallest energy that a particle can have under the given conditions, i.e. the energy of the lowest quantized state. If the box is sufficiently large, this energy, corresponding to the smallest frequency possible for standing light waves, is so small that it can be neglected. Putting $W = h\nu$ and $g = h\nu/c$, we thus get

$$dN = \frac{8\pi V}{c^3} \frac{\nu^2\, d\nu}{e^{h\nu/kT}-1}. \tag{186 a}$$

The product $dN\, h\nu$ is the *energy* of the radiation in the spectral frequency interval $d\nu$. This energy referred to unit frequency interval and unit volume is called the 'spectral energy density'. Denoting it by ρ_ν we have $dN\, h\nu = V\rho_\nu\, d\nu$, and hence

$$\rho_\nu = \frac{8\pi}{c^3} \frac{h\nu^3}{e^{h\nu/kT}-1}. \tag{186 b}$$

This formula, discovered by Planck in the year 1900, was the starting-point of the development of the quantum theory.†

We shall now show that, in the special case of a photon gas, the assumption $W_0 = 0$, upon which the preceding formulae are based, remains valid irrespective of the temperature.

† As a matter of fact Planck's quantum theory arose as the outcome of his efforts to *interpret* equation (186 b) which he discovered semi-empirically.

The general method of determining the parameter W_0, or the equivalent parameter λ, is to solve the equation

$$N = \frac{8\pi V}{h^3} \int_0^\infty \frac{g^2\, dg}{e^{\lambda+\mu W}-1}$$

where N is regarded as a given constant quantity *independent of the temperature*, i.e. of the parameter μ. This makes it necessary and possible to determine λ as a definite function of the temperature, for, since μ depends on the temperature, λ must also depend on the temperature if N is to be constant.

So long as we are dealing with a system of material particles, described either by symmetrical or antisymmetrical functions, their number must not change with change of temperature if we are to continue to discuss the *same* system. When, however, we regard radiation as a system of photons, it is no longer necessary to regard their number as constant and this number becomes irrelevant. This is because photons, unlike real material particles, can be created or annihilated by emission or absorption by the walls of the box. Thus the walls must not be supposed to be totally reflecting if we wish to follow the change of radiation with change of temperature. The permanence of the number N of photons with change of temperature is therefore neither possible nor necessary, and it is replaced by the permanence of the parameter λ. But at low temperatures we have seen that $\lambda = 0$, and if λ is constant then it must be equal to zero for all temperatures. This result can be derived more directly as follows: The parameter λ was introduced in equation (139) and still earlier in equation (133) because the 'state sum' $Z = \sum_\alpha e^{-\mu E_\alpha}$ could not be evaluated for a system consisting of a definite number of particles, described either by symmetrical or antisymmetrical functions, and so had to be replaced by the sum $\sum_{N,E} e^{-\lambda N-\mu E}$ involving a similar probability law both for the values of the energy E and the number of particles N. Now for radiation the number of particles is irrelevant, and the sum $Z = \sum e^{-\mu E_\alpha}$ can be evaluated directly. This is obviously equivalent to putting $\lambda = 0$ in the general equations obtained from evaluating $\sum e^{-\lambda N-\mu E}$. The fact that the number of particles is meaningless when radiation is regarded as a system of photons can be considered as another indication of the superficial and limited character of the analogy between radiation and matter. It shows that the photons are much more similar to the phonons or 'sound quanta' which have been considered in the preceding

section. As has been already stated, this analogy serves to undermine the current belief in the reality of photons rather than to support it.

In fact, the main difference between the photons and the obviously fictitious phonons is that the optical spectrum extends to infinity, whereas the acoustical spectrum is limited by a certain maximum wave number of the order of magnitude of the reciprocal distance between the nearest atoms in the corresponding solid body.

It should be mentioned that Planck's formula (186 b) can be derived in exactly the same way as the formula (185) for the heat energy of a solid. To obtain it, we need only drop the first term in (185) (which corresponds to longitudinal vibrations) and replace the velocity of transverse sound waves c_t by the velocity of light c. The upper limit of the integral ν_{0t} must of course be replaced by infinity to obtain the total energy of radiation.

The derivation of the distribution law (186), given in § 28, is based on the idea of mutual action between the separate particles of the 'gas' and hence, strictly speaking, does not apply to a photon gas, because photons exert no action whatever on one another. It is, however, possible, as pointed out by W. Bothe, to obtain this law by considering the statistical equilibrium of a mixture of radiation with some other gas, e.g. an electron gas, whose particles may be supposed to interact with the photons. This interaction or 'collision' of free electrons with photons is manifested by a scattering of the latter, as in the Compton effect.

Let us consider a group of $Q_{i'}$ electrons in a set of states i' and a group of $Q^*_{k'}$ photons in a set of states k'. We can define the number of transitions, due to their mutual action, into states of the sets i'' and k'' respectively, by the expression

$$f_{(i'k')\to(i''k'')} Q_{i'} Q^*_{k'} (g_{i''} - Q_{i''})(g^*_{k''} + Q^*_{k''}).$$

Equating this to the number of reverse transitions and cancelling the coefficients f, we find the equilibrium condition

$$\frac{Q_{i'}}{g_{i'} - Q_{i'}} \cdot \frac{Q^*_{k'}}{g^*_{k'} + Q^*_{k'}} = \frac{Q_{i''}}{g_{i''} - Q_{i''}} \cdot \frac{Q^*_{k''}}{g^*_{k''} + Q^*_{k''}},$$

which, in conjunction with the energy equation

$$W_{i'} + W^*_{k'} = W_{i''} + W^*_{k''},$$

leads to the distribution laws

$$\frac{Q_i}{g_i - Q_i} = A e^{-\mu W_i}, \qquad \frac{Q^*_k}{g^*_k + Q^*_k} = A^* e^{-\mu W^*_k},$$

the second of which is equivalent to Planck's law (with due allowance for the indefiniteness of the number of photons N*).

In view of its importance and theoretical interest, we shall give one more derivation of Planck's law—the derivation given by Einstein in 1917. It is based upon the consideration of the interaction between a photon gas and an atom or a number of similar atoms, the interaction in question consisting not in the *scattering* of the photons (as by the electron gas) but in their emission or absorption by the separate atoms (or by the bound electrons). The translatory motion of the atoms is of secondary importance, the transitions connected with the emission or absorption of light being essentially transitions between different 'internal' stationary states of an atom *supposed to be at rest.*

We may assume that the number of atoms is small enough to justify the application of Boltzmann's law to determine their distribution among the different 'internal' stationary states, irrespective of the character—symmetrical or antisymmetrical—of the wave functions which should be used for an exact description of the whole system of atoms.

Let N_k be the (relative or absolute) number of atoms in a stationary state k, and let us examine the number of transitions per unit time between two states k' and k'' accompanied by absorption or emission of radiation in which the atoms are 'bathed'.

If the energy of the state k'' ($W_{k''}$) is larger than that of k' ($W_{k'}$), then transitions $k' \to k''$ involving the absorption of light will be due exclusively to the action of the latter, whereas the reverse transitions $k'' \to k'$ involving the emission of light will be partially forced ('induced') and partially spontaneous. Since the probabilities of the forced transitions $k' \rightleftharpoons k''$ must be proportional to the spectral density of the radiation for the resonance frequency $\nu = \nu_{k'k''} = (W_{k''} - W_{k'})/h$, they can be written in the form $B_{k'k''}\rho_\nu$ and $B_{k''k'}\rho_\nu$ respectively. The probability of a spontaneous transition $k'' \to k'$, on the other hand, will be a constant $A_{k''k'}$ independent of ρ_ν. The equilibrium condition for a balance between the total number of direct and reverse transitions $k' \rightleftharpoons k''$ is expressed by the equation

$$N_{k'} B_{k'k''} \rho_\nu = N_{k''}(B_{k''k'}\rho_\nu + A_{k''k'}). \tag{187}$$

Now as we have shown in § 18, the coefficients $B_{k'k''}$ and $B_{k''k'}$ must be equal ('reversibility law'), so that solving (187) for ρ_ν we get

$$\rho_\nu = \frac{A_{k''k'}/B_{k''k'}}{N_{k'}/N_{k''} - 1}. \tag{187 a}$$

But according to (103 a), § 18,

$$\frac{A_{k''k'}}{B_{k''k'}} = \frac{8\pi h\nu^3_{k''k'}}{c^3}, \tag{187 b}$$

and according to Boltzmann's distribution law and Bohr's frequency condition

$$\frac{N_{k'}}{N_{k''}} = \frac{e^{-W_{k'}/\theta}}{e^{-W_{k''}/\theta}} = e^{(W_{k''}-W_{k'})/\theta} = e^{h\nu_{k''k'}/\theta}.$$

Substitution in (187 a) gives

$$\rho_{\nu_{k''k'}} = \frac{8\pi}{c^3}\frac{h\nu^3_{k''k'}}{e^{h\nu_{k''k'}/\theta}-1},$$

which is Planck's formula.

It should be mentioned that Boltzmann's law can be derived from the condition that the ratio $N_{k'}/N_{k''}$ must be a function of the frequency $\nu_{k''k'}$ alone. This condition is necessary if ρ_ν, as given by (187 a), is to be independent of the nature of the atoms present and is to depend merely on ν and the temperature. Since $\nu_{k''k'} = (W_{k''}-W_{k'})/h$, the condition reduces to the functional equation

$$\frac{N_{k'}}{N_{k''}} = f(W_{k''}-W_{k'}), \tag{187 c}$$

whose solution is $N_k = \text{const. } e^{-W_k/\theta}$.

The preceding considerations are incomplete in two respects, namely, in regard to the translatory motion of the atoms and in regard to the structure of the radiation. They can, however, be easily extended to take account of both.

Every act of absorption or emission alters the momentum $\mathfrak{g}$ and the kinetic energy ϵ of the corresponding atom. If the momentum and the energy of the incident (and subsequently absorbed) photon, or of the emitted photon are $\mathfrak{g}^*$ and ϵ^* respectively, then by the conservation laws:

$$\mathfrak{g}''-\mathfrak{g}' = \pm\mathfrak{g}^* \tag{188}$$

$$\epsilon''-\epsilon'+W_{k''}-W_{k'} = \pm\epsilon^* \tag{188 a}$$

where the single prime indicates the initial and the double prime the final state, while the upper sign refers to absorption and the lower sign to emission. Replacing ϵ^* in the last equation by $h\nu^*$ we get the frequency condition in its correct form—a form which accounts not only for the change of internal energy, but also for the change of translational energy (the atom being no longer supposed to remain at rest):

$$\nu^* = |\nu_{k''k'}| \pm \frac{\epsilon''-\epsilon'}{h}. \tag{188 b}$$

Putting $\epsilon = g^2/2m$, we have

$$\epsilon''-\epsilon' = \frac{1}{2m}(\mathfrak{g}''+\mathfrak{g}')\cdot(\mathfrak{g}''-\mathfrak{g}') = \frac{1}{2m}(\mathfrak{g}''+\mathfrak{g}')\cdot\mathfrak{g}^* = \frac{1}{m}\mathfrak{g}\cdot\mathfrak{g}^*,$$

where $\mathfrak{g}$ denotes the average value of the momentum of the atom before and after the absorption or the emission of the photon. If the direction of this average momentum, i.e. the average direction of motion, makes an angle θ with $\mathfrak{g}^*$, the direction of the absorbed or emitted light, then

$$\epsilon''-\epsilon' = \frac{1}{m}gg^*\cos\theta,$$

or, since $g/m = v$ is the (average) velocity of the atom, and $g^* = h\nu^*/c$,

$$\epsilon''-\epsilon' = h\nu^*\frac{v}{c}\cos\theta.$$

Substituting this in (188 b) and solving for ν^*—the frequency of the absorbed or emitted light—we get

$$\nu^* = |\nu_{k''k'}|\Big/\left(1\mp\frac{v}{c}\cos\theta\right),$$

or approximately

$$\nu^* = |\nu_{k''k'}|\left(1\pm\frac{v}{c}\cos\theta\right). \qquad (188\,c)$$

We thus arrive at Doppler's well-known equation—an equation expressing the change in the frequency of the light due to the motion of the material particle absorbing or emitting it. A systematic study of the changes in the momentum and kinetic energy of the atoms connected with each act of absorption or emission of a photon, determined by equations (188) and (188 a), leads to Maxwell's (or Boltzmann's) distribution law for the kinetic energies of the atoms as a consequence of Planck's law of radiation, that is, of the distribution of energy among the photons. In particular it can easily be shown that the average kinetic energy of each atom is equal to $\frac{3}{2}\theta$. This result was established by Einstein, in 1917, in the same paper in which he introduced, for the first time, the probability description of transitions by means of the coefficients A and B. It was the first direct support of Einstein's corpuscular theory of light, for it necessitated the assumption that the emission of light (as evidenced by the 'recoil' of the atom) is a process just as unidirectional as absorption.

Every emission or absorption of a photon changes the composition of the photon gas, or, in other words, the structure of the radiation field. A simple extension of Einstein's theory in this direction—an

extension which specifies the equilibrium conditions for the radiation—has been made by W. Bothe. This extension is based on the idea that, in the process of forced emission of light, the emitted photon, being in exactly the same state as the photon which caused its emission, gets associated with the latter in what could be called a 'photon molecule'. As a matter of fact such a photon-molecule corresponds to a quantized wave system or a quantized harmonic oscillator in a state containing two or more quanta of energy, the number of these quanta being equal to the number of photons constituting the 'molecule' in question.

In a detailed analysis of the processes of absorption and emission it is therefore necessary to replace the separate photons by photon-molecules, the act of absorption consisting in a *unit* decrease and that of emission in a *unit* increase in the number of photons forming the molecule. The principle of detailed balancing requires that the number of photon-molecules of each 'size', i.e. the number N_i of photons in each state i, should remain constant in spite of the above changes. It can easily be shown that this requirement is actually fulfilled.

INDEX TO PART I

Absorption probability, 142.
angular quantum number, 94.
antisymmetric functions, 173.
antisymmetry principle, 182.
asymptotic solution, 79, 86.
axial quantum number, 94.

Bloch, F., 265, 271.
Bohr, N., 53, 87, 96.
Boltzmann's law, 194, 196, 276.
Born, M., 33, 36, 172
Bose-Einstein statistics, 198.
Bothe, W., 274, 278.
Bragg's formula, 25, 122.
Brillouin, L., 209, 263.
de Broglie, 20, 23, 55.

Characteristic values, 87.
compressibility (of metals), 221.
Compton effect, 48.
Condon, E., 105.
configuration space, 167.
contact potential difference, 244.
crystal lattice, motion of electron in a, 121, 227, 230.
— —, energy levels in a, 230, 233.

Damped vibrations, 105.
Darwin, C. G., 46.
Davisson, 25.
Debye, P., 160.
decay constant, 105.
degeneracy, 83, 96, 129, 176.
Dempster, 27.
Dirac, 142, 153, 155, 160, 182, 198.
Doppler effect, 277.
Drude, 253.

'Eigenvibrations', 64.
Einstein, A., 6, 9–12, 36, 135, 147, 198, 277.
electrical conductivity of metals, 254.
electromagnetic waves, 18, 103, 188, 190.
electron gas in metals, 215.
emission probability, 135.
exchange degeneracy, 176.
exclusion principle, 162, 182.

Fermi, E., 198, 235.
Fizeau, 7, 23, 27, 30.
fluctuations (in a gas), 197, 205.
forced transitions, 139.
Fourier's theorem, 41.
Fowler, R. H., 113, 243.
frequency relation, 53, 131.

Gamow, G., 105, 106.
Germer, 25.
Goudsmit, 152.
group velocity, 29, 230.
Gurney, 105.

Harmonic oscillator, 78–84, 160.
heat capacity of electron gas, 218.
Heisenberg, W. (uncertainty relation), 46, 47–52.
Hermitian relation, 137.
holes, Dirac's, 155.
Huygens, 5.
hydrogen-like atom, 84, 108.

Identity principle, 158.
image force, 240.

Kronig, 123.

Laplace's equation, 92.
leaking, 117.
Liouville's theorem, 43, 184.
Lorentz, 8, 256.

Magnetic susceptibility of electron gas, 220.
matrix, 136.
matrix multiplication law, 149.
Maxwell, 7, 253.
mean free path of electron, 254, 257, 261.
momentum-energy vector, 11.
multiplicity (of energy levels), 83.

Newton, 1, 5, 23, 34.
nodes, 57, 81, 88.
Nordheim, 113, 243.
normalizing condition, 126.

Orthogonality relation, 130.
overtones, 57.

Pauli, W., 152, 162, 181, 198, 201, 220.
Pauli-Fermi-Dirac statistics, 198, 215.
Peierls, 265, 271.
permutations, 176.
perturbation energy, 141.
phase-space, 42, 60, 158, 200.
phonons, 265.
photons, 12, 13, 135, 187, 272.

Planck, M., 13, 160, 272.
Poisson's equation, 163, 234.
polarization of light, 150.
potential staircase method, 100.
principal quantum number, 94.
probability amplitudes, 62.
— conception, 32.
— current, 69.
— density, 32.
— packet, 38.
— theory, classical, 61.
proper time, 10.

Quanta of energy, 13.
quantized states, 53.
— waves, 161.
quantitative quantization, 162.
quantum numbers, 94.

Radiation force, 138.
Raman effect, 269.
rectification of contacts, 250.
reflection partial, 5, 68.
— total, 5, 79, 86.
refractive index, 3.
relative motion of two particles, 171.
relativistic Schrödinger equation, 77.
relativity, 7, 9.
rest-energy, 11.
reversibility law, 71, 144, 196, 207.
rotator, 98.
Rupp, 26.

Scattering of electrons, 258, 268.
scattering of photons, 275.
Schrödinger, E., 30, 64, 75, 77, 140, 170.
Schrödinger's equation, for stationary states, 74.
Schrödinger's equation, generalized, 140.
— —, relativistic, 77.
— —, for two particles, 170.
Schrödinger's theory of light emission, 132.
selective reflection of electrons by a crystal lattice, 122.
self-consistent field, 234.
Sommerfeld, A., 54, 224, 242, 256.
spherical harmonics, 92.
spin, 152, 153, 183.
spontaneous transitions, 139.
statistical equilibrium, 199.
symmetrical wave functions, 173.
superposition principle, 61, 124.

Tamm, Ig., 124.
tensor of energy, 18.
thermionic current, 242.
Thomas-Fermi equation, 235.
Thomson, G. P., 26.
transition probability, 133.
transmission coefficient, 69.
transposition, 177.
tunnel effect, 112.
Tyndall effect, 258.

Uhlenbeck, 152.
uncertainty relation, 40, 59.

Velocity, imaginary, 67.
—, group, 29, 230.

Wave equation of Schrödinger (see Schrödinger's equation).
wave number, 15.
waves, standing, 56.
Wiedemann-Franz law, 255.